Errata for "Discretization Methods for Control Systems Design"

Perry Anthony Blackmore

- pg. 12 (equation 2.2.2): right brace missing after second $\mathcal{R}(s)$
- pg. 13 (equation 2.2.3): $\mathbb{N}$ over summation should be N
- pg. 14 (last paragraph and henceforth): the notation $\mathcal{P}(s) \in \mathbb{C}^{n\times m}$ is to be understood as the space of functions which map s into the space $\mathbb{C}^{n\times m}$
- pg. 21 (equation 3.2.18): definition is obviously only valid for square systems
- pg. 23 (equation 3.2.27): $H_T^v(s)$ should be $H_T^v(z)$
- pg. 25 (last equation): $g(\sigma)$ preferable to $g(s)$
- pg. 25 (last paragraph): αT^{-1} should be $(1-\alpha)T^{-1}$
- pg. 29 (first sentence): remove "strictly"
- pg. 30 (equation set 3.4.13): these equation naturally apply for all $\boldsymbol{\beta}$ such that $\det(\mathbf{I}_n - \boldsymbol{\beta}\tilde{\mathbf{A}}) \neq 0$
- pg. 32 (last paragraph): it should be added that, as stated in Chapter 2, the optimization is for a given input signal
- pg. 34 (Lemma 3.5): for consistency of notation we should have

$$f^{-1}(z) = \frac{z-1}{(T-\alpha)z+\alpha}$$

- pg. 38 (equation 3.5.30): space missing between equations (between $\mathbf{0}$ and $\mathbf{A}^T$)
- pg. 38 (last two sentences): should read "That is, the sum of the singular values squared is increased. Conversely, as α is decreased, the sum of the singular values squared is decreased."
- pg. 42 (point 2): We have attempted in this thesis to isolate the factors which affect the magnitude of discretization errors—the Hankel singular values are one of these factors. One of the problems we faced was in trying to vary one of the parameters which affect discretization error, while holding others constant (which also have a relationship to discretization error). Despite this difficulty, the vast amount of simulation work that was undertaken, demonstrated to a large degree that the magnitude of the Hankel singular values significantly influence the magnitude of the discretization errors produced, provided they are viewed in conjunction with other factors. The statement on pg. 95 which states that the factors identified in the open-loop are not so conclusively demonstrable in the closed-loop setting is mainly due to the difficulty in isolating the factors.
- pg. 45 (2 paragraphs following equation 3.6.12): Naturally equation 3.6.12 is not applicable for the $\epsilon = 0$ and $\epsilon = 1$ case. The phenomena in the limiting situation is being addressed in this section.
- pg. 47 (second last equation): $d\tau$ should be dt

- pg. 52 (Proposition 3.1 and following): The proposition that the discretization process can be modelled as adding a white noise source to each analog integrator clearly has some limitations. Due to the presence of memory, the errors generated are more lowpass in nature. Also a "sufficiently rich" input signal can never be perfectly achieved due to aliasing considerations. However the arguments presented in this section are parallel arguments to those used in the analysis of the propagation of quantization errors in finite wordlength applications. In the analysis of quantization errors, the whiteness assumption is never actually obtained. Despite this, useful practical results can still be obtained. This fact provides the motivation and rationale of the analysis of this section. Simulation results confirm the usefulness of the results obtained.

- pg. 75-77: $\mathbf{T}^{\sim}(z)$ represents the reciprocal transpose of $\mathbf{T}(z)$ and for consistency of notation should be replaced by $\mathbf{T}^{-}(z)$

- pg. 75-77: $\Pi_{-}[\cdot]$ and $\Pi_{+}[\cdot]$ are the stable and antistable projection operators respectively and for consistency of notation should be replaccd by $[\cdot]_{-}$ and $[\cdot]_{+}$ respectively

- pg. 95 (third paragraph): the 3 occurrences of k should be replaced by N

- pg. 98 (sentence after equation 4.5.8): sentence should read, "The initial degradation........"

- pg. 114 (last sentence before Section 5.3): While objections may be raised against choosing long sampling periods for continuous-time discretization methods, this thesis has attempted to show that good performance can still be possible in this case. Our motivation in selecting long sampling periods was basically twofold. As is well known, good performance can always be obtained (providing numerical problems are not introduced) if the sampling period is low enough. Hence it makes little sense to compare different discretization methods with rapid sampling. Secondly and more importantly, there are applications where a short sampling period is not possible due to limitations in processing power. By still being able to use the continuous method of discretization in these situations, the design is likely to benefit (see arguments in Section 1.2).

- pg. 151 (proof of Lemma 3.10): the occurrences of α should be replaced by a_1, and β by a_2

DISCRETIZATION METHODS FOR CONTROL SYSTEMS DESIGN

Note to Examiners

The programs described in Appendix F of the thesis can be obtained via anonymous ftp on faceng.anu.edu.au (id number 150.203.43.3).

The following procedure is required.

1. "ftp faceng.anu.edu.au" with the username **anonymous**, and the password given by your user id.
2. "cd /pub"
3. "bin"
4. "get perry.tar.Z"
5. "mkdir discretization"
6. move perry.tar.Z to discretization
7. "cd discretization"
8. "zcat perry.tar.Z | tar xf -"

If you require any assistance, please feel free to email me at perry@faceng.anu.edu.au.

Discretization Methods for Control Systems Design

Perry Anthony Blackmore

B.Sc. (University of Melbourne)
Grad. Dip. Ed. (Melbourne College of Advanced Education)
Grad. Dip. Sc. (Australian National University)

April 1995

A thesis submitted for the degree of Doctor of Philosophy
of The Australian National University

Department of Engineering
Faculty of Engineering and Information Technology
The Australian National University

Statement of Originality

These doctoral studies were conducted under the supervision of Dr Iven M.Y. Mareels and Prof. Darrell Williamson, with Dr Robert Bitmead acting as advisor.

The contents of this thesis are the results of original research, performed in coöperation with Dr Iven M.Y. Mareels and Prof. Darrell Williamson. Approximately 70% of the work is my own. The research described in this thesis has not been submitted for any other degree or award in any other university or educational institution.

Sections of this work have been presented at conferences and submitted to journals, as detailed below.

[1] P.A. Blackmore and I.M.Y Mareels. Convex optimization techniques for analog controller discretization. submitted to *IEEE Transactions on Automatic Control* in April 1995.

[2] P.A. Blackmore, D. Williamson and I.M.Y Mareels. Discrete-time simulation techniques for continuous-time systems. submitted to *Automatica* in April 1995.

[3] P.A. Blackmore, D. Williamson and I.M.Y Mareels. Discretization of analog controllers by integral approximation. submitted to *International Journal of Control* in April 1995.

[4] P.A. Blackmore, D. Williamson and I.M.Y Mareels. Signal invariant transformation techniques for the discretization of analog controllers. submitted to *IEEE Transactions on Automatic Control* in April 1995.

[5] P.A. Blackmore, I.M.Y. Mareels, and D. Williamson. Survey of closed-loop controller discretization techniques with application to a two-tank apparatus. submitted to *Automatica* in April 1995.

[6] P.A. Blackmore and D. Williamson. Optimal discretization of analog controllers. in *Proceedings of the European Control Conference*, pages 1703–1708, Groningen, The Netherlands, 1993.

[7] P.A. Blackmore, I.M.Y. Mareels, and D. Williamson. Signal invariant transformation techniques for the discretization of analog controllers. in *Proceedings of the 33rd IEEE Conference on Decision and Control*, pages 249–254, Orlando, Florida, 1994.

[8] P.A. Blackmore and I.M.Y. Mareels. Convex optimization techniques for analog controller discretization. submitted to *34th IEEE Conference on Decision and Control* in February 1995.

During my doctoral studies, I have also worked in areas not directly related to the focus of this thesis. This work is contained in the following papers.

[9] P.A. Blackmore and R.R. Bitmead. Duality between the discrete-time Kalman filter and LQ control law. to appear in *IEEE Transactions on Automatic Control.* accepted December 1994.

[10] P.A. Blackmore and I.M.Y. Mareels. FIR filter design via convex optimization. to be submitted to *IEEE Transactions on Signal Processing*

PA Blackmore

Perry A. Blackmore
April 1995

Acknowledgements

I would like to thank my supervisor, Iven Mareels, for his efforts over the last three years. I thank him for his encouragement and guidance which has extended to areas beyond those pertaining to the academic. Thanks for being a great person. I thank Darrell Williamson for his drive in motivating much of the work of this thesis. I also thank Michael Green and Robert Bitmead for the help they have given me. A special thanks goes David Clements at the University of New South Wales for providing a great amount of assistance with the implementational aspects of Chapter 5.

To all the people in the Department of Engineering I thank you for your friendships, especially the guys in the office. To Werner, I thank you for the endless philosophical debates and the ferocious table tennis battles. To my Christian brothers—'Nuth, Aidan, Jeremy, Bob, and Brenan—I thank you for your encouragement and support.

To Bert and Carole Gibson and many of the members of the Westminster Presbyterian Church I express my gratitude for your ministry, your efforts, and your convictions.

The deepest gratitude goes to my parents. They raised me in manner that encouraged me to think. They gave me an opportunity to make it to this point. I love them both very much.

Of course to my beloved Kim, thank you for marrying me. Just how do you put up with my idiosyncrasies?

To the LORD—I now know that you are holy and I am not.

“Holy, holy, holy, is the LORD of hosts: the whole earth is full of his glory.”

Isaiah 6:3

Abstract

Given the widespread use of digital computers in the analysis, design and implementation of modern control systems, there exists a need for effective and practical discretization methods. Motivated by this need, this thesis develops new discretization techniques for use in control systems analysis and design, as well as in implementation. Both open-loop and closed-loop methods of discretization are considered. While there exist other discretization methods that are commonly in use, the analysis of this thesis demonstrates that the new methods enjoy many advantages.

Open-loop methods applicable to control system simulation, filter design and feedforward control design are developed. Factors which affect the generation and propagation of discretization errors are identified by analytical, heuristic and experimental arguments. An algorithm for open-loop discretization is presented which takes these factors into account. The fundamental idea of this method is the replacement of the analog integrators of a continuous-time system by discrete-time approximations. This is done in such a way as to optimize a given cost function with respect to a given input. The motivation of this work is to develop accurate discretization methods while limiting the complexity of the resulting discretized system. The work results in a better understanding of the discretization process in the control systems context. Connections are made between discretization and concepts from control systems analysis.

Closed-loop discretization methods are developed for the digital re-design of analog controllers. Three methods are presented which are based upon the theories of signal invariant transformations, optimal control, and convex optimization. The re-design methods which are developed exhibit particular advantages over existing methods, and together form a powerful range of techniques for the designer.

An extensive survey of existing re-design methods found in the literature is undertaken. Comparisons between these methods and those developed in this thesis are made from an analytical viewpoint as well as from a practical and an implementational viewpoint. The efficacy of the schemes developed in this thesis is demonstrated.

Contents

Statement of Originality i

Acknowledgements iii

Abstract v

Nomenclature xii

1 Introduction **1**

1.1 The Open-Loop Problem 2

1.2 The Closed-Loop Problem 5

1.3 Thesis Structure 7

1.4 Contribution of this Thesis 10

2 Problem Statement **11**

2.1 Introduction 11

2.2 Open-Loop Problem Statement 11

2.3 Closed-Loop Problem Statement 13

3 Open-Loop Problem **16**

3.1 Introduction 16

3.2 Signal Invariant Transformations 17

3.2.1 Concept 17

3.2.2 Theory 18

3.3 Motivation for the Open-Loop Discretization Scheme 25

3.4 Problem Formulation 28

3.4.1 First Order Case 28

3.4.2 Second Order Case 30

3.4.3 Mixed Order Case 32

3.4.4 The Open-Loop Problem Statement 32

3.5 Special Cases to Gain Insight into the Problem 33

3.5.1 Special Case $\bar{\mathbf{A}} = \mathbf{0}$ 33

3.5.2 Special Case $\bar{\mathbf{A}} = \mathbf{A}$ 39

3.5.3 Special Case $\bar{\mathbf{A}} = \{a_{ij}\delta_{ij}\}$ 40

3.6 The Effects of Approximation Order on Discretization Error 41

3.6.1 Scalar Result 42

3.6.2 Second Order Result 43

3.6.3 Comparing Discrete Systems Approach 45

3.6.4 Summary of Factors which Affect Discretization Error 48

3.6.5 Selection Criteria for the Allocation of Approximation Order . . 50

3.7 The Effects of State Space Structure on Discretization Error 51

3.7.1 An Excursion Into Operational Amplifiers 51

3.7.2 Optimal State Space Structure for the Minimization of Integrator Noise 53

3.8 The Discretization Algorithm 56

3.9 Simulation Results 59

3.9.1 Examination of the Effects of Bandwidth upon Discretization Error 60

3.9.2 Examination of the Effects of the Hankel Singular Values upon Discretization Error 63

3.9.3 Examination of the Effects of State Space Structure upon Discretization Error . . . . : 64

3.9.4 Case Study 67

3.10 Conclusion 69

4 Closed-Loop Problem **70**

4.1 Introduction 70

4.2 H_2 Optimal Control Method 71

4.2.1 Introduction 71

4.2.2 Primary Optimization - l_2 Minimization of Discretization Error at an Intersample Point 72

4.2.3 Secondary Optimization and Main Algorithm 80

4.3 Convex Optimization Method 86

4.3.1 Introduction 86

4.3.2 Problem Formulation 87

4.3.3 The Use of "qdes" 90

4.3.4 Unstable $\mathcal{R}(s)$ 92

4.4 Integral Approximation Method 93

4.4.1 Introduction 93

4.4.2 Problem Formulation 94

4.4.3 The Algorithm 94

4.5 Simulation Studies 95

4.5.1 H_2 Optimal Control Method 96

4.5.2 Convex Optimization Method 101

4.5.3 Integral Approximation Method 102

4.6 Conclusions 106

5 Evaluation of Closed-Loop Methods **108**

5.1 Introduction 108

5.2 The Apparatus 110

5.2.1 Identification 110

5.2.2 Analog Controller Design 112

5.3 Algorithm Descriptions 114

5.3.1 Parameter Selections 118

5.4 Results 120

5.5 Analysis 132

5.6 Conclusions 134

6 Conclusions 136

6.1 Overview of Thesis 136

6.1.1 The Open-Loop Problem 136

6.1.2 The Closed-Loop Problem 138

6.2 Further Work 139

Appendices 141

A Factorization Theory 141

B Stability Theory for Sampled-Data Systems 143

C Sampled-Data Frequency Response Methods 147

D Derivation of Error Bounds 149

D.1 Proof of Lemma 3.9 149

D.2 Proof of Lemma 3.10 150

D.3 Proof of Lemma 3.11 151

E The Continuous-Discrete Lyapunov Equation 155

F Algorithms Developed 160

F.1 Algorithms of Chapter 3 160

F.2 Algorithms of Chapter 4 162

F.3 Algorithms of Chapter 5 164

Nomenclature

Sets and Spaces

$\mathbb{R}$, $\mathbb{C}$	fields of real and complex numbers
$\mathbb{R}^{n\times m}$, $\mathbb{C}^{n\times m}$	spaces of $n \times m$ real and complex constant matrices
$\mathbb{Z}^+$	set of non-negative integers
$\mathcal{C}^0([0,\infty),\mathbb{C}^n)$	vector space of piecewise-continuous functions from $[0,\infty)$ to $\mathbb{C}^n$ that are bounded on compact subsets of $[0,\infty)$ and are continuous from the left at every point except the origin
$\mathcal{S}(\mathbb{Z}^+,\mathbb{C}^n)$	space of sequences from $\mathbb{Z}^+$ to $\mathbb{C}^n$
$\mathcal{L}_p[0,\infty)$	Lebesgue space of measurable functions $f(t)$ from $[0,\infty)$ to $\mathbb{C}^n$ which have finite L_p norm, i.e. $\lVert f\rVert_p \triangleq (\int_0^\infty \sum_{i=1}^n \lvert f_i(t)\rvert^p \, dt)^{\frac{1}{p}} < \infty$
$l_p(\mathbb{Z}^+)$	space of sequences $\mathbf{v}(k)$ from $\mathbb{Z}^+$ to $\mathbb{C}^n$ which have finite l_p norm (definition of l_p norm below)
H^∞	Hardy space, functions bounded and analytic outside the open unit disc
$H^\infty_\perp$	complementary part of H^∞, bounded and analytic within the unit disc
RH^∞, $RH^\infty_\perp$	real-rational subspaces of H^∞ and $H^\infty_\perp$ respectively

Notation and Operators

a, $\mathbf{a}$, $\mathbf{A}$	scalar, column vector, and matrix respectively
$\mathbf{A}^T$, $\mathbf{A}^*$, $\mathbf{A}^\#$	transpose, complex conjugate transpose and pseudo-inverse of a constant matrix $\mathbf{A}$, respectively
$\mathrm{tr}(\mathbf{A})$	trace of a $n \times n$ matrix
$\mathbf{I}_n$	$n \times n$ identity matrix
$\mathbf{0}$	arbitrary size zero matrix
$\mathrm{diag}(\mu_1,\ldots,\mu_n)$	$n \times n$ matrix with diagonal elements $(\mu_1,\ldots,\mu_n)$ and all other elements zero
$\lambda_j(\mathbf{A})$	eigenvalues of a matrix $\mathbf{A}$
$\pi(\mathbf{A})$	number of eigenvalues of $\mathbf{A}$ in the open right half-plane
$\nu(\mathbf{A})$	number of eigenvalues of $\mathbf{A}$ in the open left half-plane
$\mathfrak{R}(\mathbf{A}), \mathfrak{N}(\mathbf{A})$	range space and null space of a matrix $\mathbf{A}$
$f(\cdot)$, $F(\cdot)$	scalar functions
$\mathbf{f}(\cdot)$, $\mathbf{F}(\cdot)$	vector or matrix values functions

$\mathbf{P}(s)$, $\mathbf{P}(z)$	continuous-time and discrete-time transfer function matrices respectively
$\delta(P)$	relative degree of a scalar rational transfer function P; difference between the number of poles and the number of finite zeros
$\mathbf{P}^{-}(z)$	the reciprocal transpose $(\mathbf{P}(z^{-1}))^T$
$[\mathbf{P}(z)]_{-}$, $[\mathbf{P}(z)]_{+}$	stable and antistable projection of $\mathbf{P}(z)$ respectively i.e. if $\mathbf{P}(z) \in RH^{\infty} \cup RH^{\infty}_{\perp}$ then $\mathbf{P}(z)$ can be written as $\mathbf{P}(z) = [\mathbf{P}(z)]_{-} + [\mathbf{P}(z)]_{+}$ where $[\mathbf{P}(z)]_{-} \in RH^{\infty}$, $[\mathbf{P}(z)]_{+} \in RH^{\infty}_{\perp}$
$\sigma_i(\mathbf{P}(e^{j\theta}))$	frequency dependent singular values of $\mathbf{P}(\cdot)$
$\bar{\sigma}(\mathbf{P}(e^{j\theta}))$	maximum Hankel singular value of $\mathbf{P}(\cdot)$
$\left[\begin{array}{c\|c} \mathbf{A} & \mathbf{B} \\ \hline \mathbf{C} & \mathbf{D} \end{array}\right]$ [A,B,C,D]	state space realization of a rational transfer matrix $\mathbf{P}(\cdot) = \mathbf{C}(s\mathbf{I}_n - \mathbf{A})^{-1}\mathbf{B} + \mathbf{D}$ (also define $[\mathbf{A},\mathbf{B},\mathbf{C}] \triangleq [\mathbf{A},\mathbf{B},\mathbf{C},\mathbf{0}]$)

Discrete-Time Norms (where defined)

Scalar Sequences

$$\|v\|_p \triangleq \left(\sum_{k=0}^{\infty} |v(k)|^p\right)^{\frac{1}{p}} \quad 1 \leq p < \infty$$

$$\|v\|_{\infty} \triangleq \sup_{k\geq 0} |v(k)|$$

Vector Sequences

$$\|\mathbf{v}\|_p \triangleq \left(\sum_{i=1}^{n} \|v_i\|_p^p\right)^{\frac{1}{p}} \quad \mathbf{v}(k) \in \mathcal{S}(\mathbb{Z}^+, \mathbb{C}^n)$$

where v_i is the $i'th$ component of $\mathbf{v}(\cdot)$

$$\|\mathbf{v}\|_{\infty} \triangleq \max_{1\leq i\leq n} \|v_i\|_{\infty}$$

Transfer Function Matrix

$$\|\mathbf{P}\|_2 \triangleq \left[\int_0^{2\pi} \sum_{i=1}^{p} \left[\sigma_i\left(\mathbf{P}(e^{j\theta})\right)\right]^2 \, d\theta\right]^{\frac{1}{2}}$$

if $\mathbf{P}(z) \in \mathbb{C}^{n\times m}$ then $p = \min\{m, n\}$

$$= \left[\mathrm{tr}\left(\sum_{k=0}^{\infty} \mathbf{h}(k)\mathbf{h}^T(k)\right)\right]^{\frac{1}{2}}$$

where $\mathbf{h}(k)$ is the impulse response of $\mathbf{P}(z)$

$$\|\mathbf{P}\|_{\infty} \triangleq \sup_{0\leq\theta<2\pi} \bar{\sigma}\left(\mathbf{P}(e^{j\theta})\right)$$

Chapter 1

Introduction

Automatic control has played an important role in the advance of engineering and science. In our modern society, control systems affect everyone's life. For instance, in modern houses, the temperature and humidity are automatically regulated for comfortable living; in modern aviation, feedback control systems are employed for safe and comfortable air travel; in industry, control systems increase the efficiency of production; and in defence, automatic control has application in many diverse areas ranging from missile guidance to vibration control in helicopters.

Feedback control systems can be classified in a variety of ways. One classification of particular importance to this thesis involves the nature of the signals flowing in the control systems—systems can be classified as continuous-time, discrete-time or sampled-data. Many control systems occurring in nature are continuous-time—they have internal signals that evolve continuously in time with control signals that flow in the system continuously without interruption. Many of the human body's regulatory systems can be described as continuous-time control systems. Conversely, a discrete-time control systems have signals that evolve discretely in time. Many economic systems can be described in this framework. Systems having both discrete and continuous signals are called hybrid systems or sampled-data systems. These systems generally arise when analog signals of a continuous-time system are sampled for control purposes, with the control signals calculated via a digital computer. This type of system is a major focus of this thesis.

The digital computer has changed the methodology of control systems implementation.

In the past, continuous-time controllers were the norm in industry. Typically they were realized electrically—using resistors, capacitors, inductors, and operational amplifiers—or mechanically—using hydraulics and pneumatics. While continuous-time controllers are still used in practice, applications of digital control are becoming more widespread. Digital computers are now widely used in the direct on-line control of processes. This shift can be attributed to the advances made in microelectronics and digital signal processing in recent decades. Typically, digital controllers are more reliable, cheaper, more compact, less susceptible to ageing and have improved sensitivity to noise and parameter variations.

The digital computer has also impinged on the techniques of controller design. Many of the classical tools of control systems design—such as Bode analysis, Nyquist analysis, and root-locus analysis—can be more effectively utilized using the digital computer. Design techniques which were formally very time consuming can now be rapidly computed using the digital computer. Many of the burdens of controller design have been eliminated. In short, the digital computer has greatly influenced the theory and practice of control engineering.

Of course the digital computer has limitations and introduces new problems. In the area of controller implementation, limitations in computing speed may introduce unacceptable time delays. Finite wordlength of the digital processor can lead to problems such as the introduction of unstable limit cycles. In control systems design, inaccuracies in system simulation result from discretization errors and finite wordlength effects.

The goal of the work presented in this thesis is to develop techniques to minimize discretization errors in systems design and controller implementation—in both open-loop and closed-loop. These problems are now discussed in greater detail.

1.1 The Open-Loop Problem

There are a number of instances in control systems design where a continuous-time system is discretized in isolation, or more specifically when a block or element is discretized without considering the interaction with other elements. In this thesis, this shall be referred to as **open-loop discretization**. Some examples of applications of open-loop discretization are:

Filter Design: The traditional approach to infinite impulse response (IIR) discrete-time filter design involves first the design of a continuous-time filter and then transformation into a discrete-time filter (c.f. [62]). There are a number of reasons for this approach. For example, the art of continuous-time IIR filter design is highly advanced and so it is advantageous to incorporate this approach. Also many useful continuous-time IIR methods have relatively simple closed-form design formulae. It is clear that an effective means of discretization is essential for the success of this approach.

Feedforward Controller Design: There are examples in control system design where discrete-time elements are designed in isolation—e.g. the design of discrete-time feedforward elements. A typical approach is to first design an acceptable analog element and then perform open-loop discretization.

Simulation: A major application of open-loop discretization is the simulation of a system containing analog elements on a digital computer. Obviously, this is integrally related to control system design. During the design process, the designer is typically concerned with how the given system responds to different inputs—steps, ramps, and sinusoids for example. As the controller design is refined, a number of simulations may be carried out for for a given controller. As a result, the simulation procedure is carried out many times. The importance of accurate and effective simulation procedures is clearly seen. In essence, the simulation of continuous-time dynamical systems generally requires the numerical solution of a set of differential equations. Discretization or approximation is a feature of these techniques.

For the discretization of linear, time-invariant dynamical systems where a transfer function description of the system exists, one can think of the discretization process as transforming a continuous-time transfer function into a discrete-time one. Naturally, there is no unique equivalence between continuous-time and discrete-time systems for all input signals—with any discretization method there is a loss of information. As a result of discretization, errors are generated and propagated—a natural concern is to minimize these errors. Before deciding on a particular method of discretization, the designer must decide which properties of the continuous-time system are important to preserve. There is a wide range of possible frequency-domain or time-domain properties. Furthermore, the complexity of the discretization method is important. Discretization methods which produce systems of order ten times that of the continuous prototype are unacceptable in many applications. Clear trade-offs exist.

There exists a large body of literature dealing with discretizing continuous-time systems and in particular, simulating continuous-time systems on a digital computer. Some of the early methods are:

Picard's method, which generates a series of functions which converge to the true solution. This method is of theoretical interest, and is really only of practical value when an analytic solution is available. Alternative methods are obviously preferable.

Taylor series methods, which apply a recursion formula to generate a series of solution points to the differential equation. This is a good method when analytic expressions for the high order derivatives are known.

Runge-Kutta methods, which are derived from the Taylor series methods. Runge-Kutta methods are widely used, having many advantages of the Taylor series methods, but avoiding the problems inherent with computing high order derivatives.

Predictor-corrector methods, which are a family of methods involving a pair of formulae, the predictor and the corrector. These methods enable tight control over the error at each point by the repeated application of the corrector and the computation times are generally shorter than those of the Runge-Kutta methods. A disadvantage of the predictor-corrector methods is that they are dependent upon another method to generate a set of starting values.

Many texts are available which discuss the above methods, see for example [44].

In the survey paper of Kowalczuk [50], two broad classes of open-loop discretization methods are identified. The most popular are the **closed-form transformations**, in which the discretizing transformation is chosen to minimize the difference between the response of the digital model and the the response of the analog prototype at the sampling points $\{t_k = kT, \; k \;$ an integer$\}$. The error depends on the input signal applied. As a result, distinctions are made according to whether the spectrum of the input signal is bandlimited to the Nyquist range $(\omega \; : \; \omega < \omega_N = \pi/T)$ or not.

Closed-form approximations can be further divided into **direct transformations** and **indirect transformations**. Examples of the former include signal invariant transformation methods [79, 82, 84], forming element methods [11, 22, 39, 69, 47, 48, 49, 70, 79, 81], and convolution approximation methods [78, 80]. A feature of direct transformation methods is that the prototype system $\mathcal{F}(s)$ is treated as a whole, and an attempt is made to preserve the input-output relationship. The indirect approach involves treating $\mathcal{F}(s)$ as a number of sub-systems such as integrators or differentiators. These sub-systems are replaced by discrete-time operators. Examples of indirect transformations include mapping of differentials [42, 62, 82, 84], forward and backward Euler approximations [42, 62, 69, 82, 84], bilinear transformations [11, 42, 62, 64, 76, 82, 84], and matched z-transforms [42, 62, 84].

The second broad class of discretization methods can be referred to as **open-form approximations** [4, 35, 62, 63, 64, 70]. These methods generally involve some form of iterative optimization algorithm and do not result in closed form solutions. Most of these methods use a frequency based criterion and are quite often computationally expensive. In engineering, perhaps the most well-known of these methods is that developed by Deczky [24, 62, 63], in which an L_p frequency-domain criterion is minimized.

In this thesis, a unified theory is presented which addresses essential aspects of the open-loop discretization procedure. Unlike many discretization schemes presented in the literature, the procedure developed attempts to first determine the magnitude of the expected discretization errors, and then discretize with an appropriate complexity. The proposed algorithm gives the designer control over the order of the discrete-time system. This enables the design of filters with low complexity which still retain essential properties of the prototype system.

Attempts are made, where possible, to examine the discretization process from an engineering perspective. For example, a number of relationships are found between the bound on the magnitude of the discretization error and the Hankel singular values. To our knowledge, this approach is a new contribution.

1.2 The Closed-Loop Problem

The design and implementation of sampled-data controllers motivates the study of minimizing discretization error in a closed-loop setting. Of course there are many practical issues in the design of sampled-data controllers—the design of the digital-to-analog converters, analog-to-digital converters, sample-and-hold elements, and anti-aliasing filters to mention a few. However the design of the control algorithm used in the computer is the focus in this thesis.

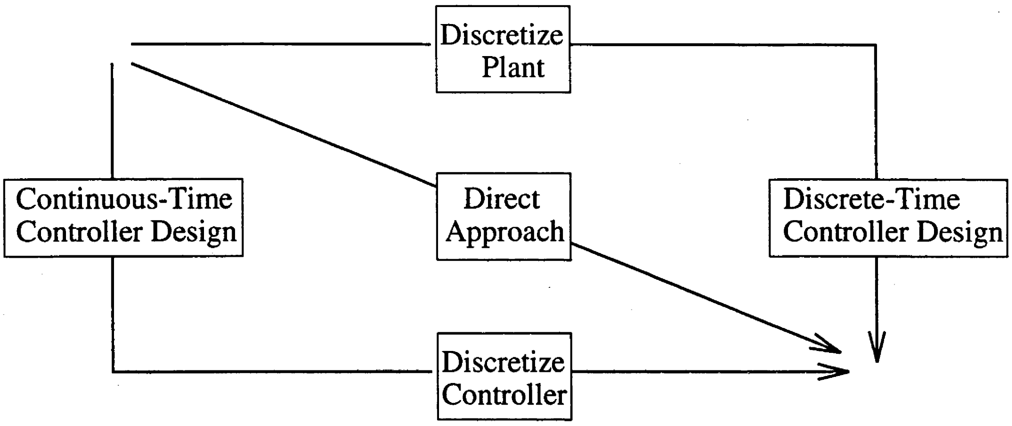
Discretize Plant
Continuous-Time Controller Design
Direct Approach
Discrete-Time Controller Design
Discretize Controller

Figure 1-1: Methods of sampled-data controller design

Figure 1-1 represents the three choices of sampled-data design method. These are:

Discretize Plant Approach: The given model of the continuous-time plant is replaced by an equivalent discrete-time model. A discrete-time design method (such as those found in [6, 32]) is then used to generate a discrete-time controller. There are two main disadvantages with this approach. First, *a priori* decisions such as the choice of sampling-rate and possible sampling-skew must be made. But sensible choices of these quantities depend upon the performance of the final closed-loop system. If these quantities are not estimated correctly, the whole design may have to be repeated. A second problem of this approach is that intersample behaviour is ignored at the start of the design. For this reason, the resulting performance of the digital controller may be poor.

Discretize Controller Approach: A continuous-time controller may be designed based on the continuous-time plant. The continuous-time controller can then be discretized using a number of different methods. Controller discretization methods can be found in [1, 32, 43, 45, 50, 52, 53, 54, 58, 60, 61, 65, 66, 67, 71, 85]. This approach is favourable if the designer is more familiar with continuous-time design. Loop shaping problems and constraint satisfaction can be achieved more effectively than with direct methods. Additionally, in contrast to the "Discretize Plant Approach", this approach allows intersample behaviour to be taken into account in the design.

Direct Approach: Methods have recently been developed for the direct design of sampled-data controllers. These have the advantage of incorporating the true continuous-time specifications. Advances in this area can be found in [9, 10, 17,

> 18, 19, 20, 21, 26, 38, 40, 41, 46, 73, 74, 75, 86]. In particular, the sampled-data H_2 problem has been tackled by Khargonekar and Sivashankar [46], Bamieh and Pearson [9], Chen and Francis [18, 20] and Chen [17]. The sampled-data H_∞ problem has been addressed by Toivonen [75], Bamieh and Pearson [10], and Kabamba and Hara [41]. Dullerud and Francis have developed theory for the L_1 sampled-data problem. Despite the mathematical difficulties present in describing sampled-data systems, the above work represents a significant advance. However, the calculations are generally very forbidding. Experience indicates that these methods do not allow the designer a great deal of flexibility—there are many situations in which a better design can be achieved by "indirect" approaches, particularly if the designer has more experience in the area of continuous-time controller design or if there are multiple control objectives. These approaches are also difficult to apply to complex systems. Furthermore, the H_2 sampled-data approach often yields digital controllers that generate large control signals. This can be a problem in practical settings.

There are many situations in which the "Discretize Controller Approach" is preferable, producing a superior performance to the "Direct Approach". It is therefore important that effective controller discretization schemes exist. This thesis presents optimization based controller discretization algorithms which are shown to produce sampled-data controllers with excellent performance. The algorithms developed are shown to have many advantages over other discretization schemes presented in the literature.

1.3 Thesis Structure

A brief outline of the progression of ideas in this thesis is as follows:

Signal Invariant Transformations

After the problems addressed in this thesis are formulated at a conceptual level in Chapter 2, the introductory material of Chapter 3 presents a review of the existing theory of signal invariant transformations. This material is foundational for much of the theory presented in this thesis. The signal invariant transformation, for a particular input signal, allows a continuous-time system to be discretized so that the sampled output of the continuous-time system is equivalent to the output of the discrete-time system in response to the sampled input. The signal invariant transformation of a continuous-time system has an order typically equal to the order of the continuous-time prototype plus the order of the continuous-time model whose impulse response generates the input signal.

Discretization via Integral Approximation

The open-loop discretization method is introduced in Sections 3.3 and 3.4. The basis of the scheme is a modification of the Newton-Cotes formulae for numerical integration. Each integrator of an n-dimensional system is replaced by a Newton-Cotes type approximation. The order of the approximation of each integrator may vary. The method is an open-form approximation technique using iterative optimization. A time-domain criterion is used with a formulation in terms of state space structures. Optimization is performed with respect to a given signal. This method gives a lower order discrete-time system than the signal invariant transformation technique without a severe loss of time-domain performance.

Factors which affect discretization error are then identified. These factors are incorporated into the discretization methodology and a unified theory is presented. An analysis of the effects of state space structure is given and theory is presented which enables the selection of state space structures favourable for minimizing discretization error.

Controller Discretization Using H_2 Optimal Control Theory

Chapter 4 presents the closed-loop discretization methods. Existing literature approaches this problem in basically two ways:

1. Open-loop analog-to-digital conversions using, for example, the bilinear, forward difference, backward difference, step invariant and impulse invariant transformations [6, 42, 84].

2. Closed-loop analog-to-digital design using signal invariant transformations [84] or other techniques [1, 43, 45, 65].

It is now well recognised that, when designing a digital controller to replace an analog controller, the closed-loop properties of the continuous-time system (consisting of analog controller and plant) should be taken into consideration. Consequently, closed-loop discretization generally results in better performance and as such is preferred, particularly when the sampling rate is low. In many cases, it is possible to reduce the sampling rate by a factor of 10 in comparison to the open-loop approaches. With this motivation, this thesis presents three methods of closed-loop controller discretization. The first method, found in Section 4.2, uses an extension of the open-loop signal invariant transformation theory. Optimal control theory is used and a number of algorithms are proposed. The method is shown to perform favourably when compared with existing algorithms. A disadvantage of this method is that digital controllers of high order are produced. The sampled-data controller reduction method of [56] is shown to effectively deal with this problem.

Controller Discretization Using Convex Optimization Techniques

In Section 4.3, a similar problem is solved using signal invariant transformations and convex optimization techniques. This theory is based on the work of [13]. The convex optimization technique is shown to be an extremely powerful method of design. A variety of constraints can be incorporated into the design which makes this the most flexible method of controller discretization available. Again the disadvantage of this method is that the resulting digital controllers are of high order.

Controller Discretization Using Integral Approximations

The open-loop discretization method developed in Chapter 3 is extended to the closed-loop setting in Section 4.4. Although a non-linear optimization problem results, this method has the advantage that model order reduction techniques are not required in the final phase of design. Unfortunately the method suffers from the existence of local minima. However a simple technique which greatly alleviates the numerical difficulties associated with this problem is presented.

Comparison of Discretization Methods

The controller discretization algorithms developed in this thesis are then compared to existing methods presented in the literature. The comparison involves a short simulation study in Section 4.5 and an extensive analysis in Chapter 5. The analysis in Chapter 5 is based on computer simulations, practical experimentation and subjective analysis. The three methods developed in this thesis are demonstrated to be extremely effective in solving the controller discretization problem.

1.4 Contribution of this Thesis

The preceeding discussion demonstrates a four-fold contribution of this thesis:

Open-Loop Discretization Method

- comprehensive treatment of the open-loop discretization problem
- theory for replacing the integrators of a continuous-time system by discrete-time approximants, which allows the designer to control the complexity of the discretization
- techniques which enable the designer to identify the factors which affect discretization error
- theory which gives good state space structures for discretization

Closed-Loop Discretization Methods

- discretization method using signal invariant transformations and optimal control theory
- discretization method based on signal invariant transformations and convex optimization
- discretization method based on integral approximations

Survey of Discretization Methods

- comprehensive survey of closed-loop discretization methods

Algorithms

- open-loop discretization—includes algorithms for parameter optimization and optimization of state space structure for discretization
- controller discretization using H_2 optimal control theory
- controller discretization using convex optimization
- controller discretization based on integral approximations
- sampled-data controller reduction

Chapter 2

Problem Statement

2.1 Introduction

As stated in Chapter 1, the primary focus of this thesis is to develop discretization techniques for use in the open-loop and closed-loop settings. This chapter refines this statement by presenting formal problem statements for both problems. The problems are formulated at a conceptual level at this point, and are further elaborated and refined in Chapters 3 and 4.

2.2 Open-Loop Problem Statement

The discretization problems considered in this thesis are restricted to linear, finite dimensional systems. A continuous-time system can be viewed as a mapping from $\mathcal{C}^0([0,\infty),\mathbb{C}^n) \rightarrow \mathcal{C}^0([0,\infty),\mathbb{C}^n)$, whereas a discrete-time system can be viewed as a mapping from $\mathcal{S}(\mathbb{Z}^+,\mathbb{C}^n) \rightarrow \mathcal{S}(\mathbb{Z}^+,\mathbb{C}^n)$. The general problem of discretization is finding a transformation $\mathfrak{D}$ which maps a given continuous-time system $\boldsymbol{\mathcal{G}}(s)$ into a discrete-time system $\mathbf{G}(z)$, i.e.

$$\mathbf{G}(z) = \mathfrak{D}\left(\boldsymbol{\mathcal{G}}(s)\right) \tag{2.2.1}$$

Ideally the transformation would be such that the responses of both systems would be identical (at the sampling instances) for any input signal, not necessarily piecewise constant. However this is not possible in reality when realizability and complexity issues are encountered. Only some of the properties of $\mathcal{G}(s)$ can be preserved under the $\mathfrak{D}$-transformation. Therefore, the discretization problem can be thought of as finding a $\mathfrak{D}$ which minimizes

$$\mathfrak{F}_o\{\mathcal{G}(s) - \mathbf{G}(z)\}$$

for some distance function $\mathfrak{F}_o$.

Assume that the continuous-time system $\mathcal{G}(s)$ is driven by a signal $\mathbf{r}(t)$ for $t \geq 0$. Assume also that the discrete-time system is driven by the sampled $\mathbf{r}(t)$ which is denoted by $\mathbf{S}_T\mathbf{r}(t)$. The sampling operator $\mathbf{S}_T$ (with period T) is defined in Appendix B. Let the continuous-time input signal $\mathbf{r}(t)$ be generated by the impulse response of a strictly proper system $\mathcal{R}(s)$. With a slight abuse of notation, let the sampled $\mathbf{r}(t)$ be represented by $\mathbf{S}_T\mathcal{R}(s)$. Similarly, let the sampled output of $\mathcal{G}(s)$ be represented by $\mathbf{S}_T\mathcal{G}(s)\mathcal{R}(s)$.

The minimization of

$$\mathfrak{F}_o\{\mathbf{S}_T\mathcal{G}(s)\mathcal{R}(s) - \mathbf{G}(z)\mathbf{S}_T\mathcal{R}(s)\}$$

is sought for some distance function $\mathfrak{F}_o$ which is now defined.

Let $\mathcal{G}(s)$ represent a strictly proper, linear time-invariant system. Let $\mathcal{R}(s)$ represent a strictly proper, linear time-invariant reference model whose impulse response is $\mathbf{r}(t),\ t \geq 0$. Let the response of $\mathcal{G}(s)$ (with zero initial conditions) to $\mathbf{r}(t)$ be represented by $\mathbf{y}(t),\ t \geq 0$. Furthermore, suppose $\mathbf{y}(kT)$ is the sampled values of $\mathbf{y}(t)$ at $t = kT$ with $k = 0, 1, \ldots, \infty$ for some sampling period $T > 0$. Similarly, let $\mathbf{r}(kT)$ be the sampled $\mathbf{r}(t)$ for $k = 0, 1, \ldots, \infty$. Let $\mathbf{G}(z)$ be restricted to the class of linear, time-invariant systems. Finally let $\hat{\mathbf{y}}(kT)$ be the response of $\mathbf{G}(z)$ (with zero initial conditions) to $\mathbf{r}(kT)$.

Define the distance function $\mathfrak{F}_o$ to take the form of a cost function $\mathcal{J}_N$ where

$$\mathfrak{F}_o\{\mathbf{S}_T\mathcal{G}(s)\mathcal{R}(s) - \mathbf{G}(z)\mathbf{S}_T\mathcal{R}(s) \triangleq \mathcal{J}_N \triangleq \sum_{k=0}^{N} \|\mathbf{y}(kT) - \hat{\mathbf{y}}(kT)\|_2^2 \qquad (2.2.2)$$

This cost function is chosen for several reasons. It gives a good measure of error between the continuous-time and discrete-time systems. It is computationally easy to calculate. By varying N, the discretization error during the transient response (by making NT small) or the steady-state (by making NT large) can be targeted. As $N \to \infty$, with T fixed, $\mathcal{J}_N$ is the l_2 norm—in this case Parseval's theorem indicates that $\mathcal{J}_N$ is related to the frequency-domain error.

A transformation $\mathfrak{D}$ is sought to minimize (2.2.2) subject to the realizability of $\mathbf{G}(z)$. This is still a very broad problem and it does not address the problem of the complexity of $\mathbf{G}(z)$. Therefore a parameterization of $\mathbf{G}(z)$ is established so that restrictions on the complexity of $\mathbf{G}(z)$ are managed. Represent this parameterization as $\mathbf{G}(z, \mathbf{p})$ for some parameter $\mathbf{p}$. Define $\hat{\mathbf{y}}_p(kT)$ to be the response of $\mathbf{G}(z, \mathbf{p})$ (with zero initial conditions) to $\mathbf{r}(kT)$.

The open-loop problem addressed by this thesis is to find a parameter $\mathbf{p}$ *which achieves the minimization*

$$\min_{\mathbf{p}} \mathcal{J}_N(\mathbf{p})$$

where

$$\mathcal{J}_N(\mathbf{p}) \triangleq \sum_{k=0}^{\mathbb{N}} \|\mathbf{y}(kT) - \hat{\mathbf{y}}_p(kT)\|_2^2 \qquad (2.2.3)$$

with

$$\mathbf{y}(t) \triangleq \boldsymbol{\mathcal{G}}(s)\mathbf{r}(t)$$
$$\hat{\mathbf{y}}(kT) \triangleq \mathbf{G}(z, \mathbf{p})\mathbf{r}(kT)$$

The choice of parameterization is presented in Chapter 3.

2.3 Closed-Loop Problem Statement

The general aim of *digital re-design* is to replace an analog controller $\boldsymbol{C}(s)$ by a *digital* controller $\mathbf{C}_d(z)$ so that the performance of the closed-loop sampled-data system 'approximates' the performance of the closed-loop analog system. The addition of a discrete element in an analog environment requires that there is a process of signal conversion so that the digital and analog components can be interfaced in the same system. Therefore analog-to-digital converters, digital-to-analog converters, sample-and-hold devices and multiplexers are inherent in the overall design. Naturally there are complex issues associated with the design and implementation of these components. However the following assumptions are made in this thesis:

- The analog-to-digital converters have infinite wordlength.
- The digital-to-analog converters are ideal: that is, there is zero acquisition time, zero aperture time, zero settling time and zero hold-mode droop.
- The hold function is a zero order hold. Although the theory of this thesis does not preclude other forms of hold functions, the zero order hold is the most common

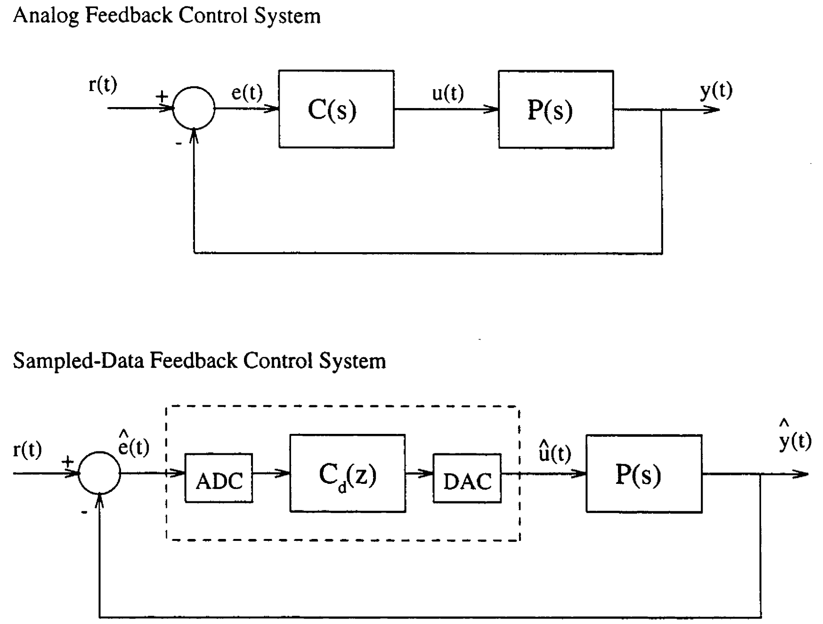
Analog Feedback Control System
r(t)
+
-
e(t)
C(s)
u(t)
P(s)
y(t)
Sampled-Data Feedback Control System
r(t)
+
-
ê(t)
ADC
C_d(z)
DAC
û(t)
P(s)
ŷ(t)

Figure 2-1: Analog and digital control

in practical applications.

These assumptions are common to most discretization methods and are reasonable for most applications.

Anti-aliasing filters are also a common addition in the re-design process. These are generally analog devices placed before the samplers. They do not appear explicitly in the re-design theory of this thesis. However if the dynamics of these filters have a significant effect on the overall system dynamics, they can be readily accounted for with the theory developed.

In Figure 2-1, an *analog* unity feedback control system is illustrated. The system consists of a strictly proper, linear time-invariant plant $\mathcal{P}(s) \in \mathbb{C}^{n\times m}$ and a proper linear-time invariant controller $\mathcal{C}(s) \in \mathbb{C}^{m\times n}$. Assume also a strictly proper linear time-invariant reference model $\mathcal{R}(s) \in \mathbb{C}^{n\times 1}$, whose impulse response generates the reference signal $\mathbf{r}(t)$. The output of the analog closed-loop system is represented as $\mathbf{y}(t)$. The sampled-data unity feedback control system shows the digital controller $\mathbf{C}_d(z)$ with analog-to-digital converter and digital-to-analog converter. The output of the sampled-data system is represented as $\hat{\mathbf{y}}(t)$.

Assumptions: *Suppose the analog controller $\boldsymbol{\mathcal{C}}(s)$ has been chosen such that:*

A The closed-loop system is asymptotically stable.

B The output tracking error $\|\mathbf{y}(t) - \mathbf{r}(t)\| \to 0$ as $t \to \infty$ where $\mathbf{r}(t) = \mathcal{L}^{-1}\{\boldsymbol{\mathcal{R}}(s)\}$ is the impulse response of the reference model $\boldsymbol{\mathcal{R}}(s)$.

Let the plant, continuous-time controller, and reference model be considered as operators, mapping $\mathcal{C}^0([0,\infty),\mathbb{C}^n) \to \mathcal{C}^0([0,\infty),\mathbb{C}^n)$—denote these operators as simply $\boldsymbol{\mathcal{P}}$, $\boldsymbol{\mathcal{C}}$ and $\boldsymbol{\mathcal{R}}$ respectively.

The closed-loop operator of the analog closed-loop system is given by

$$\boldsymbol{\mathcal{H}} : \mathcal{C}^0([0,\infty),\mathbb{C}^n) \to \mathcal{C}^0([0,\infty),\mathbb{C}^n) \text{ where } \boldsymbol{\mathcal{H}} = \boldsymbol{\mathcal{P}}\boldsymbol{\mathcal{C}}(\mathbf{I}_n + \boldsymbol{\mathcal{P}}\boldsymbol{\mathcal{C}})^{-1} \tag{2.3.1}$$

The digital controller $\mathbf{C}_d$ is an operator mapping $\mathcal{S}(\mathbb{Z}^+,\mathbb{C}^n) \to \mathcal{S}(\mathbb{Z}^+,\mathbb{C}^n)$.

The closed-loop operator of the sampled-data closed-loop system is given by

$$\mathfrak{H}_T = \boldsymbol{\mathcal{P}}\mathbf{H}_T\mathbf{C}_d\mathbf{S}_T(\mathbf{I}_n + \boldsymbol{\mathcal{P}}\mathbf{H}_T\mathbf{C}_d\mathcal{S}_T)^{-1} \tag{2.3.2}$$

where the *sampling operator* $\mathbf{S}_T$, and the *zero-order hold operator* $\mathbf{H}_T$ are defined in Appendix B. The operator $\mathfrak{H}_T$ maps $\mathcal{C}^0([0,\infty),\mathbb{C}^n) \to \mathcal{C}^0([0,\infty),\mathbb{C}^n)$.

The closed-loop problem addressed by this thesis is to find a stabilizing digital controller $\mathbf{C}_d(z)$ which minimizes

$$\mathfrak{F}_c\{(\mathfrak{H}_T - \boldsymbol{\mathcal{H}})\boldsymbol{\mathcal{R}}\} \tag{2.3.3}$$

for some distance function $\mathfrak{F}_c$.

The operator $\mathfrak{F}_c$ is algorithm dependent. Stability in this sense means exponential stability; input-output stability is not addressed by the algorithms developed. Appendix B contains a discussion of stability of sampled-data systems. For a detailed treatment, see [19, 29].

The closed-loop problem can be alternatively stated: find a stabilizing digital controller $\mathbf{C}_d(z)$, such that

$$\tilde{\mathfrak{F}}_c\{\hat{\mathbf{y}}(t) - \mathbf{y}(t)\}$$

is minimized for some distance function $\tilde{\mathfrak{F}}_c$, in response to a reference signal $\mathbf{r}(t)$.

Chapter 3

Open-Loop Problem

3.1 Introduction

This chapter presents a comprehensive treatment of the open-loop discretization problem. As stated in Chapter 1, the general aim of the discretization method developed in this thesis is to give the designer control over the complexity of the discretization. This is achieved by first identifying the factors which influence the magnitude of the discretization error and then adjusting the complexity accordingly. For example, if a large sampling period is used which is conducive to large errors, then a complex discretization method may be warranted.

In this chapter, the signal invariant transformation is used as the benchmark of open-loop discretization performance. It is also used extensively in the closed-loop techniques of Chapter 4. Because of its important role , this chapter presents a detailed treatment of signal invariant transformations. As mentioned in Chapter 1, this transformation allows a discretization which gives perfect matching in response to a reference signal. If n_1 is the order of the analog system and n_2 is the order of the model that generates the input signal, then the order of the discrete system produced by a signal invariant transformation is typically equal to $n_1 + n_2$. This complexity may be detrimental to performance in some applications, for example in an environment where there are restrictions on the computational time. A lower order discretization is generally sought in such cases. A fundamental concern of this chapter is finding a discretized system of lower order (if one exists) which allows "near perfect" matching.

The particular parameterization used in the open-loop discretization method is based on a Newton-Cotes method of integral approximation. In an n-dimensional continuous-time system there are n integrators. The discretization method replaces each of these integrators by a modified Newton-Cotes discrete-time approximation, parameterized by a vector $\mathbf{p}$. The parameterization allows the designer to control the complexity of the discretization by allowing each analog integrator to be replaced by either a zeroth, first, or second order digital approximation. An optimization is then performed to find a $\mathbf{p}$ which minimizes the open-loop cost function (2.2.3), introduced in Chapter 2. The discretization problem is formulated in terms of a state space description.

A major emphasis of this chapter is to identify the factors which contribute to and affect the magnitude of discretization errors. These factors are identified by analytical, heuristic and experimental means. As mentioned in Chapter 1, these factors are linked to concepts found in control theory, such as the Hankel singular values. The effect of state space structure on the generation and propagation of discretization errors is also investigated in this chapter.

This chapter is organized as follows. Section 3.2 presents the theory of signal invariant transformations. A motivation of the open-loop discretization scheme is given in Section 3.3. The open-loop problem that is initially formulated in Chapter 2 is fully formulated in Section 3.4. A number of special cases are looked at in Section 3.5 in order to gain an insight into the parameterization used. By looking at the effects of approximation order on discretization error, the factors which affect discretization error are then identified in Section 3.6. The effects of state space structure on discretization error are identified in Section 3.7 and an algorithm which givens an "optimal" state space structure is presented. A summary of results appears in Section 3.8 and the discretization algorithm is presented. Simulation results are presented in Section 3.9, with conclusions drawn in Section 3.10.

3.2 Signal Invariant Transformations

3.2.1 Concept

Figure 3-1 gives a diagrammatic representation of the signal invariant transformation. A continuous-time system $\mathbf{G}(s)$ with a continuous-time input and output is shown. The system $\mathbf{G}_d(z)$ is the signal invariant transformation of $\mathbf{G}(s)$ with respect to the

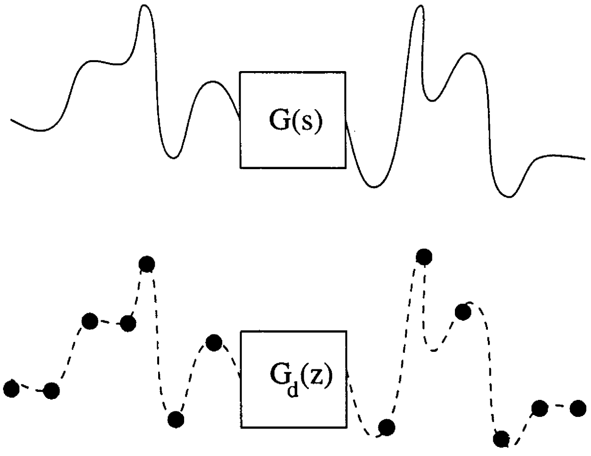

Figure 3-1: Representation of the signal invariant transformation

input signal if the sampled input applied to $\mathbf{G}_d(z)$ yields an output equivalent to the sampled output of $\mathbf{G}(s)$. This relationship is shown in the figure.

3.2.2 Theory

Consider a continuous-time transfer function matrix $\mathcal{H}(s)$ having a minimal state space representation

$$\mathcal{H}(s) = \mathbf{C}_1(s\mathbf{I}_{n_1} - \mathbf{A}_1)^{-1}\mathbf{B}_1 \tag{3.2.1}$$

In the special case where $\mathcal{H}(s)$ is single input, single output (SISO), the transfer function is denoted by

$$\mathcal{H}(s) = \mathbf{c}_1^T(s\mathbf{I}_{n_1} - \mathbf{A}_1)^{-1}\mathbf{b}_1$$

The impulse response, $\mathbf{h}(t)$, with respect to an impulse at time 0, is given by

$$\mathbf{h}(t) = \mathcal{L}^{-1}\{\mathcal{H}(s)\} = \begin{cases} \mathbf{C}_1 e^{\mathbf{A}_1 t}\mathbf{B}_1 & t \geq 0 \\ \mathbf{0} & t < 0 \end{cases} \tag{3.2.2}$$

where $\mathcal{L}$ denotes the Laplace transform operator. The z-transform $\mathbf{H}_T(z)$ of the uniformly sampled values $\{\mathbf{h}(kT);\ k \text{ non-negative integer}\}$ of $\mathbf{h}(t)$ is given by

$$\mathbf{H}_T(z) = \mathcal{Z}\{\mathbf{C}_1 e^{\mathbf{A}_1 kT}\mathbf{B}_1\} = z\mathbf{C}_1(z\mathbf{I}_{n_1} - \mathbf{F}_1)^{-1}\mathbf{B}_1 \ ; \ \mathbf{F}_1 \triangleq e^{\mathbf{A}_1 T} \tag{3.2.3}$$

Definition 3.1 *The discrete-time system $\mathbf{H}_T(z)$ in (3.2.3) is said to be the impulse invariant transformation of $\mathcal{H}(s)$, and is written*

$$\mathbf{H}_T(z) \triangleq Z_T\{\mathcal{H}(s)\} \tag{3.2.4}$$

The subscript T is used to explicitly indicate that $\mathbf{H}_T(z)$ depends on the sampling period T.

Lemma 3.1 *(i) The impulse invariant transform $\mathbf{H}_T(z)$ of $\boldsymbol{\mathcal{H}}(s)$ in (3.2.1) is a proper discrete transfer function given by*

$$\mathbf{H}_T(z) = \mathbf{C}_1\mathbf{B}_1 + \mathbf{C}_1(z\mathbf{I}_{n_1} - \mathbf{F}_1)^{-1}\mathbf{F}_1\mathbf{B}_1 \ ; \ \mathbf{F}_1 \triangleq e^{\mathbf{A}_1 T} \tag{3.2.5}$$

(ii) $\mathbf{H}_T(z)$ is strictly proper if and only if $\mathbf{C}_1\mathbf{B}_1 = 0$.
(iii) For $\mathbf{C}_1\mathbf{B}_1 \neq \mathbf{0}$, the Smith zeros of $H_T(z)$ are given by the eigenvalues of

$$\mathbf{F}_1 - \mathbf{F}_1\mathbf{B}_1(\mathbf{C}_1\mathbf{B}_1)^{-1}\mathbf{C}_1$$

In the SISO case, there are n_1 zeros of $H_T(z)$.
(iv) In the SISO case, if $\mathbf{c}_1^T\mathbf{b}_1 = 0$, then $\mathcal{H}(s)$ has relative degree of at least 2, but $H_T(z)$ has relative degree 1 for almost every sampling period T.

Proof: Parts (i)-(iii) follow by expansion of (3.2.3). Now suppose $\mathbf{c}_1^T\mathbf{b}_1 = 0$. Then $H_T(z)$ is of relative degree 1 if and only if $\mathbf{c}_1^T\mathbf{F}_1\mathbf{b}_1 \neq 0$. But $\mathbf{F}_1$ (see (3.2.5)) is an analytic function of T, and so $\mathbf{c}_1^T\mathbf{F}_1\mathbf{b}_1$ does not vanish for almost every T. This proves part (iv). ∎

The z-transform $\mathbf{H}_{T,\xi}(z)$ of the uniformly sampled values $\{\mathbf{h}(kT + \xi)$; k non-negative integer, $0 \leq \xi < T\}$ of $\mathbf{h}(t)$ is given by

$$\begin{aligned}
\mathbf{H}_{T,\xi}(z) &= Z_{T,\xi}\{\boldsymbol{\mathcal{H}}(s)\} && (3.2.6)\\
&= \mathcal{Z}\{\mathbf{C}_1 e^{\mathbf{A}_1(kT+\xi)}\mathbf{B}_1\} && (3.2.7)\\
&= \mathbf{C}_1 e^{\mathbf{A}_1\xi}\mathbf{B}_1 + \mathbf{C}_1 e^{\mathbf{A}_1\xi}(z\mathbf{I}_{n_1} - \mathbf{F}_1)^{-1}\mathbf{F}_1\mathbf{B}_1 \ ; \ \mathbf{F}_1 \triangleq e^{\mathbf{A}_1 T} && (3.2.8)
\end{aligned}$$

Note that $\mathbf{H}_{T,\xi}(z)$ is proper, but not strictly proper, for almost all values of T and ξ.

Definition 3.2 *The discrete-time system $\mathbf{H}_{T,\xi}(z)$ in (3.2.6)-(3.2.8) is said to be the ξ-offset impulse invariant transformation of $\boldsymbol{\mathcal{H}}(s)$, and is written*

$$\mathbf{H}_{T,\xi}(z) \triangleq Z_{T,\xi}\{\boldsymbol{\mathcal{H}}(s)\} \tag{3.2.9}$$

Some properties of the impulse invariant transformation of *cascaded* systems are now examined. Suppose

$$\mathcal{HG}(s) \triangleq \mathcal{H}(s)\mathcal{G}(s) \tag{3.2.10}$$

where the strictly proper $\mathcal{H}(s)$ is given by (3.2.1), and the proper system $\mathcal{G}(s)$ has a minimal state space representation

$$\mathcal{G}(s) = \mathbf{D}_2 + \mathbf{C}_2(s\mathbf{I}_{n_2} - \mathbf{A}_2)^{-1}\mathbf{B}_2 \tag{3.2.11}$$

Then a state space representation of $\mathcal{HG}(s)$ is given by

$$\mathcal{HG}(s) = \mathcal{H}(s)\mathbf{D}_2 + \mathbf{C}(s\mathbf{I}_{n_1+n_2} - \mathbf{A})^{-1}\mathbf{B} \tag{3.2.12}$$

where

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_1 & \mathbf{B}_1\mathbf{C}_2 \\ \mathbf{0} & \mathbf{A}_2 \end{bmatrix} \; ; \; \mathbf{B} = \begin{bmatrix} \mathbf{0} \\ \mathbf{B}_2 \end{bmatrix} \; ; \; \mathbf{C} = \begin{bmatrix} \mathbf{C}_1 & \mathbf{0} \end{bmatrix} \tag{3.2.13}$$

In the SISO case, the representation of $\mathcal{HG}(s)$ in (3.2.12) is minimal if and only if there are no pole-zero cancellations between $\mathcal{H}(s)$ and $\mathcal{G}(s)$.

Lemma 3.2 *The impulse invariant transformation $\mathbf{HG}_T(z)$ of $\mathcal{HG}(s)$, when $\mathcal{H}(s)$ and $\mathcal{G}(s)$ have different poles, defined by (3.2.12) and (3.2.13) is given by*

$$\mathbf{HG}_T(z) = \mathbf{H}_T(z)\mathbf{D}_2 + \mathbf{HG}_1(z) \tag{3.2.14}$$

where

$$\mathbf{HG}_1(z) \triangleq \mathbf{C}_1(z\mathbf{I}_{n_1} - \mathbf{F}_1)^{-1}\mathbf{F}_{12}[\mathbf{I}_{n_2} + (z\mathbf{I}_{n_2} - \mathbf{F}_2)^{-1}\mathbf{F}_2]\mathbf{B}_2$$

where $\mathbf{H}_T(z)$ is given by (3.2.5), and where the $n_1 \times n_2$ matrix $\mathbf{F}_{12}$ is given in terms of an $n_1 \times n_2$ matrix $\mathbf{P}_{12}$ by

$$\mathbf{F}_{12} = \mathbf{P}_{12}\mathbf{F}_2 - \mathbf{F}_1\mathbf{P}_{12} \tag{3.2.15}$$

$$\mathbf{P}_{12}\mathbf{A}_2 - \mathbf{A}_1\mathbf{P}_{12} = \mathbf{B}_1\mathbf{C}_2 \tag{3.2.16}$$

In particular, when $\mathbf{D}_2 = \mathbf{0}$, $\mathbf{HG}_T(z)$ is strictly proper.

Proof: Define the $(n_1 + n_2) \times (n_1 + n_2)$ similarity transformation

$$\mathbf{P} \triangleq \begin{bmatrix} \mathbf{I}_{n_1} & \mathbf{P}_{12} \\ \mathbf{0} & \mathbf{I}_{n_2} \end{bmatrix}$$

where $\mathbf{P}_{12}$ is defined by the equation (3.2.16). From the theory of Sylvester equations (see [7] for example), $\mathbf{P}_{12}$ is uniquely defined because by assumption $\mathbf{A_1}$ and $\mathbf{A_2}$ have different spectra. From (3.2.13)

$$\mathbf{P}^{-1}\mathbf{AP} = \begin{bmatrix} \mathbf{A}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{A}_2 \end{bmatrix}$$

Moreover, from (3.2.13), $\mathbf{CB} = \mathbf{0}$ and so from Lemma 3.1

$$\mathbf{HG}_T(z) = \mathbf{H}_T(z)\mathbf{D}_2 + \mathbf{C}(z\mathbf{I}_{n_1+n_2} - \mathbf{F})^{-1}\mathbf{FB} \ ; \ \mathbf{F} = e^{\mathbf{A}T}$$

Then

$$\mathbf{F} = \mathbf{P}e^{\mathbf{P}^{-1}\mathbf{AP}T}\mathbf{P}^{-1} = \begin{bmatrix} \mathbf{F}_1 & \mathbf{F}_{12} \\ \mathbf{0} & \mathbf{F}_2 \end{bmatrix} \ ; \ \mathbf{F}_2 \triangleq e^{\mathbf{A}_2 T}$$

where $\mathbf{F}_1$ and $\mathbf{F}_{12}$ are given by (3.2.3) and (3.2.15) respectively, which then gives (3.2.14). When $\mathbf{D}_2 = \mathbf{0}$, $\mathbf{HG}(z)$ is strictly proper. ▮

A similar result can be derived for the ξ-offset impulse invariant transformation $\mathbf{HG}_{T,\xi}(z)$ of $\mathcal{HG}(s)$. Note that even when $\mathbf{D}_2 = \mathbf{0}$, $\mathbf{HG}_{T,\xi}(z)$ is almost always proper and not strictly proper for $\xi \neq 0$.

Definition 3.3 *Consider a signal* $\mathbf{v}(t) = \mathcal{L}^{-1}\{\mathcal{V}(s)\}$. *Then the discrete-time system* $\mathbf{H}_T^{\mathbf{v}}(z)$ *written*

$$\mathbf{H}_T^{\mathbf{v}}(z) = Z_T^{\mathbf{v}}\{\mathcal{H}(s)\} \tag{3.2.17}$$

where

$$Z_T^{\mathbf{v}}\{\mathcal{H}(s)\} \triangleq Z_T\{\mathcal{H}(s)\mathcal{V}(s)\}\,[Z_T\{\mathcal{V}(s)\}]^{-1} \tag{3.2.18}$$

is said to be the signal invariant transformation *of* $\mathcal{H}(s)$ *in (3.2.1) with respect to* $\mathbf{v}(t)$.

The subscript T and superscript $\mathbf{v}$ on the operation $Z_T^{\mathbf{v}}\{\cdot\}$ in (3.2.17), (3.2.18) indicate that a signal invariant transformation depends explicitly on both the sampling period T *and* the reference signal $\mathbf{v}(t)$. The relationship in (3.2.18) says that the continuous-time convolution of $\mathbf{v}(t)$ with $\mathbf{h}(t)$ in (3.2.2), followed by sampling, is equal to the discrete-time convolution of the sampled values $\{\mathbf{v}(kT)\}$ with the sampled values $\{\mathbf{h}(kT)\}$. In the general multivariable setting, a restriction applies on the signal model $\mathcal{V}(s)$ to guarantee the existence of the inverse in (3.2.18). From Lemma 3.1, a necessary and sufficient condition for the existence of this inverse is that $\mathbf{C}_2\mathbf{B}_2$ is invertible, given

$\mathcal{V}(s) = [\mathbf{A}_2, \mathbf{B}_2, \mathbf{C}_2]$. It will be shown shortly that in the SISO case, this condition is not required.

By a slight abuse of the definition, $\mathbf{H}_T(z)$ in (3.2.3) was called an *impulse* invariant transformation. However there is actually no signal $\mathbf{v}(t)$ such that

$$\mathcal{V}(s) = \mathcal{L}\{\mathbf{v}(t)\} = \mathbf{I}_{n_2} \; ; \;\; Z_T\{\mathcal{V}(s)\} = \mathcal{Z}\{\mathbf{v}(kT)\} = \mathbf{I}_{n_2}$$

Definition 3.4 *Consider a signal* $\mathbf{v}(t) = \mathcal{L}^{-1}\{\mathcal{V}(s)\}$. *Then the discrete-time system* $\mathbf{H}^v_{T,\xi}(z)$ *written*

$$\mathbf{H}^v_{T,\xi}(z) = Z^v_{T,\xi}\{\mathcal{H}(s)\} \tag{3.2.19}$$

where

$$Z^v_{T,\xi}\{\mathcal{H}(s)\} \triangleq Z_{T,\xi}\{\mathcal{H}(s)\mathcal{V}(s)\}\,[Z_T\{\mathcal{V}(s)\}]^{-1} \tag{3.2.20}$$

is said to be the ξ-offset signal invariant transformation *of* $\mathcal{H}(s)$ *in (3.2.1) with respect to* $\mathbf{v}(t)$.

Example 3.1: A state space realization of the ξ-offset step invariant transformation of $\mathcal{H}(s)$ in (3.2.1) is given by

$$\left[e^{\mathbf{A}_1T}, \int_0^T e^{\mathbf{A}_1\tau}\mathbf{B}_1\,d\tau, \mathbf{C}_1e^{\mathbf{A}_1\xi}, \mathbf{C}_1\int_0^\xi e^{\mathbf{A}_1\tau}\mathbf{B}_1\,d\tau\right] \tag{3.2.21}$$

To conclude this treatment of signal invariant transformations, some results for SISO systems are presented.

Corollary 3.1 *(i) Suppose*

$$\mathcal{V}(s) = \frac{1}{s+\alpha}$$

and $\mathcal{H}(s) = \mathbf{c}_1^T(s\mathbf{I}_{n_1} - \mathbf{A}_1)^{-1}\mathbf{b}_1$ *has no poles at* $s = -\alpha$. *Then*

$$H^v_T(z) = \mathbf{c}_1^T(z\mathbf{I}_{n_1} - \mathbf{F}_1)^{-1}\mathbf{f}_{12} \tag{3.2.22}$$

where

$$\mathbf{f}_{12} = (\mathbf{F}_1 - e^{-\alpha T}\mathbf{I}_{n_1})(\mathbf{A}_1 + \alpha\mathbf{I}_{n_1})^{-1}\mathbf{b}_1 \; ; \; \mathbf{F}_1 = e^{\mathbf{A}_1T}$$

In particular, when $\mathcal{H}(s)$ *has no poles at* $s = 0$, *the step invariant transformation* $H^s_T(z)$ *(corresponding to* $\alpha = 0$*) is given by*

$$H^s_T(z) = \mathbf{c}_1^T(z\mathbf{I}_{n_1} - \mathbf{F}_1)^{-1}(\mathbf{F}_1 - \mathbf{I}_{n_1})\mathbf{A}_1^{-1}\mathbf{b}_1 \tag{3.2.23}$$

(ii) Suppose

$$\mathcal{V}(s) = \frac{1}{(s+\alpha)^2}$$

and $\mathcal{H}(s)$ has no poles at $s = -\alpha$. Then

$$H_T^v(z) = d_1 + \mathbf{c}_1^T(z\mathbf{I}_{n_1} - \mathbf{F}_1)^{-1}\mathbf{g}_1 \tag{3.2.24}$$

where

$$\begin{aligned} d_1 &= \mathbf{c}_1^T(\mathbf{F}_1 - Te^{-\alpha T}\mathbf{A}_1 - e^{-\alpha T}(1+\alpha T)\mathbf{I}_{n_1})e^{\alpha T}T^{-1}(\mathbf{A}_1 + \alpha\mathbf{I}_{n_1})^{-2}\mathbf{b}_1 \\ \mathbf{g}_1 &= (-\alpha T\mathbf{A}_1 + \mathbf{A}_1 - \alpha^2 T\mathbf{I}_{n_1} + \alpha T\mathbf{A}_1\mathbf{F}_1 - e^{\alpha T}\mathbf{A}_1\mathbf{F}_1 + \alpha^2 T\mathbf{F}_1 - \mathbf{A}_1\mathbf{F}_1 \\ &\quad + e^{\alpha T}\mathbf{A}_1\mathbf{F}_1^2)T^{-1}(\mathbf{A}_1 + \alpha\mathbf{I}_{n_1})^{-2}\mathbf{A}_1^{-1}\mathbf{b}_1 \end{aligned}$$

In particular, when $\mathcal{H}(s)$ has no poles at $s = 0$, the ramp invariant transformation $H_T^i(z)$ of $\mathcal{H}(s)$ (corresponding to $\alpha = 0$) is given by

$$H_T^i(z) = d_1 + \mathbf{c}_1^T(z\mathbf{I}_{n_1} - \mathbf{F}_1)^{-1}\mathbf{g}_1 \tag{3.2.25}$$

where

$$\begin{aligned} d_1 &= \mathbf{c}_1^T(\mathbf{F}_1 - T\mathbf{A}_1 - \mathbf{I}_{n_1})T^{-1}\mathbf{A}_1^{-2}\mathbf{b}_1 \\ \mathbf{g}_1 &= (\mathbf{I}_{n_1} - 2\mathbf{F}_1 + \mathbf{F}_1^2)T^{-1}\mathbf{A}_1^{-2}\mathbf{b}_1 \end{aligned}$$

(iii) Suppose

$$\mathcal{V}(s) = \frac{\gamma_1}{s} + \frac{\gamma_2}{s^2} \tag{3.2.26}$$

and $\mathcal{H}(s)$ has no poles at $s = 0$. Then

$$H_T^v(s) = \frac{z}{(z-1)^2}[\gamma_1(z-1)H_T^s(z) + \gamma_2 T H_T^i(z)] \tag{3.2.27}$$

where $H_T^s(z)$ and $H_T^i(z)$ are given by (3.2.23) and (3.2.25).

Theorem 3.1 *Suppose*

$$\mathcal{H}(s) = \mathbf{c}_1^T(s\mathbf{I}_{n_1} - \mathbf{A}_1)^{-1}\mathbf{b}_1 \; ; \; \mathcal{G}(s) = \mathbf{c}_2^T(s\mathbf{I}_{n_2} - \mathbf{A}_2)^{-1}\mathbf{b}_2$$

Then:

(i) The signal invariant transformation $H_T^g(z)$ of $\mathcal{H}(s)$ with respect to the signal $g(t) = \mathcal{L}^{-1}\{\mathcal{G}(s)\}$ is given by

$$H_T^g(z) = \frac{\mathbf{c}_1^T(z\mathbf{I}_{n_1} - \mathbf{F}_1)^{-1}\mathbf{F}_{12}[\mathbf{I}_{n_2} + (z\mathbf{I}_{n_2} - \mathbf{F}_2)^{-1}\mathbf{F}_2]\mathbf{b}_2}{z\mathbf{c}_2^T(z\mathbf{I}_{n_2} - \mathbf{F}_2)^{-1}\mathbf{b}_2} \tag{3.2.28}$$

where

$$\mathbf{F}_1 = e^{\mathbf{A}_1 T} \ ; \ \mathbf{F}_2 = e^{\mathbf{A}_2 T} \tag{3.2.29}$$

and $\mathbf{F}_{12}$ *satisfies (3.2.15), (3.2.16).*

(ii) $H_T^g(z)$ *is proper with at most* $n_1 + n_2$ *poles.*

(iii) $H_T^g(z)$ *is strictly proper if* $g(0) \neq 0$ *(i.e.* $\mathbf{c}_2\mathbf{b}_2 \neq 0$*) with relative degree 1 for almost all sampling periods T. The* $n_1 + n_2$ *poles of* $H_T^g(z)$ *are then given by the* n_1 *eigenvalues of* $\mathbf{F}_1$ *and the* n_2 *eigenvalues of* $\tilde{\mathbf{F}}_2$ *where*

$$\tilde{\mathbf{F}}_2 \triangleq \mathbf{F}_2 - \mathbf{F}_2\mathbf{b}_2(\mathbf{c}_2^T\mathbf{b}_2)^{-1}\mathbf{c}_2^T \tag{3.2.30}$$

(iv) If $\mathcal{H}(s)$ *is asymptotically stable and* $\mathcal{G}(s)$ *is stable (with no poles at the origin) and of minimum phase, then* $H_T^g(z)$ *is asymptotically stable for T sufficiently large.*

(v) If both $\mathcal{H}(s)$ *and* $\mathcal{G}(s)$ *are stable and of minimum phase, then* $H_T^g(z)$ *is of minimum phase for T sufficiently large.*

Proof: The expression for $H_T^g(z)$ in (3.2.28) follows from (3.2.14) and (3.2.3). It is then evident from (3.2.28) that the poles of $H_T^g(z)$ are given by the n_1 eigenvalues of $\mathbf{F}_1$, and the n_2 zeros of $z\mathbf{c}_2^T(z\mathbf{I}_{n_2} - \mathbf{F}_2)^{-1}\mathbf{b}_2$.

When $\mathbf{c}_2^T\mathbf{b}_2 \neq 0$ (ie when $g(0) \neq 0$), the zeros of $z\mathbf{c}_2^T(z\mathbf{I}_{n_2} - \mathbf{F}_2)^{-1}\mathbf{b}_2$ are given by the eigenvalues of $\tilde{\mathbf{F}}_2$ in (3.2.30).

Parts (ii) and (iii) then follow from Lemmas 3.1 and 3.2. To prove (iv), first observe that if $\mathcal{H}(s)$ is asymptotically stable, then the eigenvalues of $\mathbf{F}_1$ are inside $|z| = 1$. Now by (3.2.23)

$$\mathbf{c}_2^T(z\mathbf{I}_{n_2} - \mathbf{F}_2)^{-1}\mathbf{b}_2 = Z_T\{\tilde{\mathcal{H}}(s)\}$$

where

$$\tilde{\mathcal{H}}(s) = \mathbf{c}_2^T(s\mathbf{I}_{n_2} - \mathbf{A}_2)^{-1}\mathbf{A}_2(\mathbf{F}_2 - \mathbf{I}_{n_2})^{-1}\mathbf{b}_2$$

That is, $\mathbf{c}_2^T(z\mathbf{I}_{n_2} - \mathbf{F}_2)^{-1}\mathbf{b}_2$ is the step invariant transformation (zero order hold equivalent) of $\tilde{\mathcal{H}}(s)$. Hence by Astrom et al. [5], all limiting zeros of $\mathbf{c}_2^T(z\mathbf{I}_{n_2} - \mathbf{F}_2)^{-1}\mathbf{b}_2$ approach $z = 0$ as $T \to \infty$ if $\tilde{\mathcal{H}}(s)$ is stable and if $\tilde{\mathcal{H}}(s)$ has no zero at $s = 0$. Furthermore, $\tilde{\mathcal{H}}(0) = 0$ if and only if $\mathbf{c}_2^T(\mathbf{F}_2 - \mathbf{I}_{n_2})^{-1}\mathbf{b}_2 \neq 0$. That is, $\tilde{\mathcal{H}}(0) \neq 0$ for almost all sampling periods T. ∎

3.3 Motivation for the Open-Loop Discretization Scheme

This section presents some basic theory and a simple example to motivate the open-loop discretization method of this thesis.

Given a C^2 scalar function $g(s)$, a sampling period or discretization size T, and some integer k, the integral

$$\int_{kT}^{kT+T} g(\sigma)\, d\sigma \tag{3.3.1}$$

can be approximated via the first order Newton-Cotes approximation

$$\frac{T}{2}[g(kT) + g(kT+T)] \tag{3.3.2}$$

This is more commonly referred to as the trapezoidal rule. Many elementary calculus books show that an error of $O(T^3)$ is associated with the above approximation. The error is also proportional to the second derivative of the function $g(s)$.

Suppose the integral is approximated by

$$\alpha g(kT) + \beta g(kT+T) \tag{3.3.3}$$

where $\alpha, \beta \in \mathbb{R}$. The basis for this approximation is the fact that by the intermediate value theorem, there exists $\bar{\alpha}, \bar{\beta}$ such that

$$\int_{kT}^{kT+T} g(s)\, ds = \bar{\alpha} g(kT) + \bar{\beta} g(kT+T)$$

It should be noted at this point that for general α and β, and $g(s)$ in some class, a meaningful error bound can not be calculated. Note that if $\bar{\alpha}$ and $\bar{\beta}$ are known *a priori*, or at least approximately known, then a significant improvement can be obtained in the integral approximation in some cases.

Clearly the trapezoidal rule is a subset of the approximation scheme corresponding to $\alpha = \beta = T/2$. Further, the two other "classical" approximations—the backward Euler and forward Euler—correspond to $\alpha = 0, \beta = T$ and $\alpha = T, \beta = 0$ respectively. Further appreciation of the role of α and β can be gained from Figure 3-2. Setting $\beta = T - \alpha$, the area under the curve between kT and $kT + T$ is approximated by area "A" plus a fraction, αT^{-1}, of area "B". The three "classical" approximations—the forward Euler, backward Euler and trapezoidal rules—correspond to different proportions of the area "B".

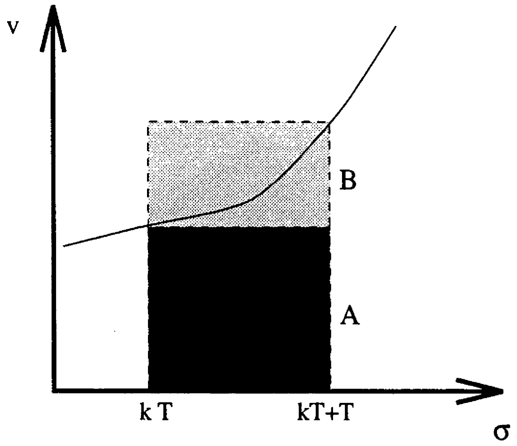

Figure 3-2: Approximation of integrals

Consider now the first order dynamical system

$$\dot{x}(t) = -ax(t) + au(t) \; ; \; u(t) = \cos(\omega t) \tag{3.3.4}$$

$$y(t) = x(t) \tag{3.3.5}$$

Equations (3.3.4) and (3.3.5) can be written in integral form, i.e.

$$y(kT+T) = y(kT) + \int_{kT}^{kT+T} -ay(\sigma) + au(\sigma)\, d\sigma \tag{3.3.6}$$

If the integral in (3.3.6) is approximated according to (3.3.3), one may ask how sensitive this approximation is to variations in α and β. To address this question, simulation results were generated using the open-loop cost function $\mathcal{J}_N(\mathbf{p})$ introduced in Chapter 2; that is

$$\mathcal{J}_N(\mathbf{p}) = \sum_{k=0}^{N} \|\mathbf{y}(kT) - \hat{\mathbf{y}}_p(kT)\|_2^2 \tag{3.3.7}$$

as a measure of discretization error.

The quantity (3.3.7) was calculated in the case $\alpha = \beta = \frac{T}{2}$ i.e. the bilinear transformation, and for α, β generated according to the optimization technique developed in Section 3.8. In both cases the quantity N was chosen to be $N = 2^{15} - 1$ (the reason for this odd number will become apparent when the algorithm is introduced in Section 3.8). The sampling period was selected to be $T = 1$ second and the quantities a and w were varied such that $0 < \omega T < \frac{1}{2}$ and $0 < aT < \frac{1}{2}$. The results are displayed in Figures 3-3 and 3-4.

Even for this simple first order system, it is evident that substantial improvement in discretization performance can be achieved through parameter optimization. This

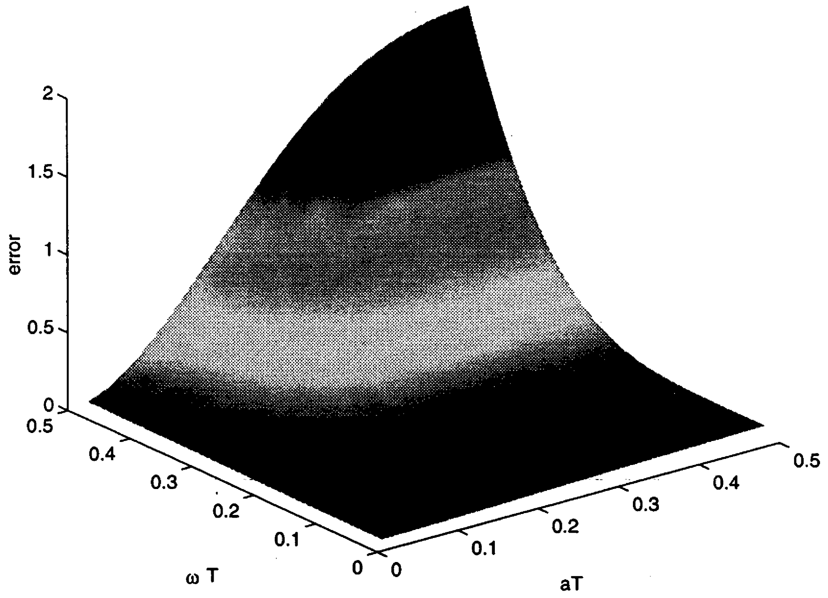

Figure 3-3: Discretization error - bilinear transformation

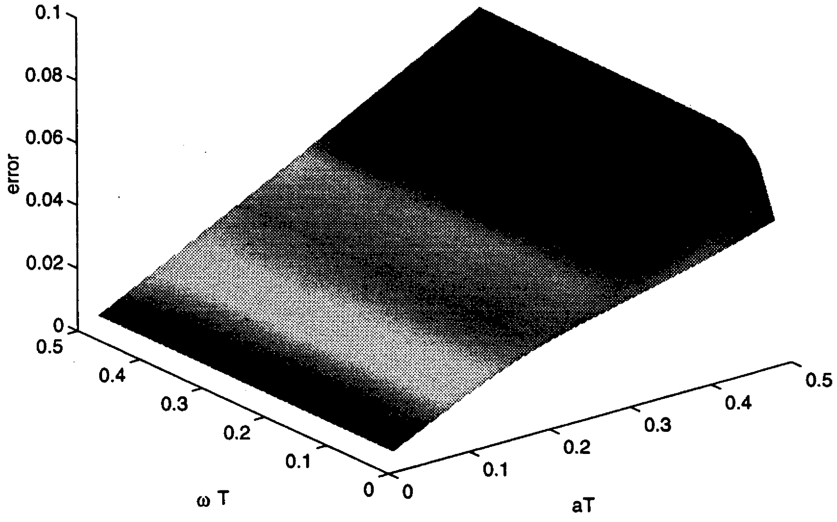

Figure 3-4: Discretization error - optimized

is particularly evident as ω and/or a become large. The importance of parameter optimization is even greater for higher order systems.

3.4 Problem Formulation

The key idea associated with the open-loop discretization scheme is the replacement of integrators of the continuous-time system by digital approximations of different orders. In this section, the basic theory is developed which enables this procedure to be performed.

3.4.1 First Order Case

In Section 3.3, the approximation of the integral

$$\int_{kT}^{kT+T} g(\sigma)\, d\sigma \tag{3.4.1}$$

was considered where $g(s)$ is a scalar function. This idea is now extended to the approximation of a continuous-time system with a **transfer function description** which is realised using a *number* of integrators. Consequently, if $\mathbf{g}(s)$ is a vector function, i.e. $\mathbf{g}(s) \triangleq [g_1(s)\ g_2(s) \ldots g_n(s)]^T$, the method generalizes to

$$\begin{aligned} \int_{kT}^{kT+T} \mathbf{g}(\sigma)\, d\sigma \quad &\approx \quad \begin{bmatrix} \alpha_1 g_1(kT) + \beta_1 g_1(kT+T) \\ \alpha_2 g_2(kT) + \beta_2 g_2(kT+T) \\ \cdots \\ \cdots \\ \alpha_n g_n(kT) + \beta_n g_n(kT+T) \end{bmatrix} && (3.4.2) \\ &= \boldsymbol{\alpha}\mathbf{g}(kT) + \boldsymbol{\beta}\mathbf{g}(kT+T) && (3.4.3) \end{aligned}$$

where

$$\boldsymbol{\alpha} \triangleq \mathrm{diag}(\alpha_1, \alpha_2, ..., \alpha_n); \quad \boldsymbol{\beta} \triangleq \mathrm{diag}(\beta_1, \beta_2, ..., \beta_n) \tag{3.4.4}$$

Suppose a strictly proper analog system $\mathcal{G}(s) = \mathbf{C}(s\mathbf{I}_n - \mathbf{A})^{-1}\mathbf{B}$ with input $\mathbf{u}$ and output $\mathbf{y}$ has a state space representation

$$\begin{aligned} \dot{\mathbf{x}}(t) &= \mathbf{A}\mathbf{x}(t) + \mathbf{B}\mathbf{u}(t) && (3.4.5) \\ \mathbf{y}(t) &= \mathbf{C}\mathbf{x}(t) && (3.4.6) \end{aligned}$$

where $\mathbf{A} = \{a_{ij}\}$, $\mathbf{B} = \{b_{ij}\}$, $\mathbf{C} = \{c_{ij}\}$. Write

$$\dot{\mathbf{x}}(t) = \bar{\mathbf{A}}\mathbf{x}(t) + \tilde{\mathbf{A}}\mathbf{x}(t) + \mathbf{B}\mathbf{u}(t) \tag{3.4.7}$$

where $\bar{\mathbf{A}}$ is a matrix with all its eigenvalues strictly in the left half plane and $\tilde{\mathbf{A}} = \mathbf{A} - \bar{\mathbf{A}}$. This gives

$$\mathbf{x}(kT+T) = e^{\bar{\mathbf{A}}T}\mathbf{x}(kT) + \int_{kT}^{kT+T} e^{\bar{\mathbf{A}}(kT+T-\sigma)}\{\tilde{\mathbf{A}}\mathbf{x}(\sigma) + \mathbf{B}\mathbf{u}(\sigma)\}\, d\sigma \tag{3.4.8}$$

The continuous-time system evolves according to (3.4.8), and discretization of the system requires numerical evaluation of the integral on the right hand side.

Using the Cayley-Hamilton theorem, one can write

$$e^{At} = \sum_{i=0}^{n-1} \Gamma_i(t)\mathbf{A}^i \tag{3.4.9}$$

where $\Gamma_i(t)$ are scalar functions of t. Therefore (3.4.8) can be written as

$$\begin{aligned}\mathbf{x}(kT+T) &= e^{\bar{\mathbf{A}}T}\mathbf{x}(kT) + \int_{kT}^{kT+T} \sum_{i=0}^{n-1} \Gamma_i(kT+T-\sigma)\bar{\mathbf{A}}^i\{\tilde{\mathbf{A}}\mathbf{x}(\sigma) + \mathbf{B}\mathbf{u}(\sigma)\}\, d\sigma \\ &= e^{\bar{\mathbf{A}}T}\mathbf{x}(kT) + \sum_{i=0}^{n-1} \bar{\mathbf{A}}^i \int_{kT}^{kT+T} \Gamma_i(kT+T-\sigma)\{\tilde{\mathbf{A}}\mathbf{x}(\sigma) + \mathbf{B}\mathbf{u}(\sigma)\}\, d\sigma\end{aligned}$$

Denote approximations to $\mathbf{x}(\cdot)$ and $\mathbf{y}(\cdot)$ by $\hat{\mathbf{x}}(\cdot)$ and $\hat{\mathbf{y}}(\cdot)$ respectively; that is

$$\hat{\mathbf{x}}(.) \approx \mathbf{x}(.), \;\; \hat{\mathbf{y}}(.) \approx \mathbf{y}(.) \tag{3.4.10}$$

An approximation to $\mathbf{x}(kT+T)$ is given by

$$\begin{aligned}\hat{\mathbf{x}}(kT+T) &= e^{\bar{\mathbf{A}}T}\hat{\mathbf{x}}(kT) + \sum_{i=0}^{n-1} \bar{\mathbf{A}}^i\Gamma_i(T)\boldsymbol{\alpha}\{\tilde{\mathbf{A}}\hat{\mathbf{x}}(kT) + \mathbf{B}\mathbf{u}(kT)\} \\ &\quad + \sum_{i=0}^{n-1} \bar{\mathbf{A}}^i\Gamma_i(0)\boldsymbol{\beta}\{\tilde{\mathbf{A}}\hat{x}(kT+T) + \mathbf{B}\mathbf{u}(kT+T)\}\end{aligned}$$

In effect, the integral corresponding to each state is approximated differently. Thus

$$\begin{aligned}\hat{\mathbf{x}}(kT+T) &= e^{\bar{\mathbf{A}}T}\hat{\mathbf{x}}(kT) + e^{\bar{\mathbf{A}}T}\boldsymbol{\alpha}\{\tilde{\mathbf{A}}\hat{\mathbf{x}}(kT) + \mathbf{B}\mathbf{u}(kT)\} \\ &\quad + \boldsymbol{\beta}\{\tilde{\mathbf{A}}\hat{\mathbf{x}}(kT+T) + \mathbf{B}\mathbf{u}(kT+T)\}\end{aligned}$$

or

$$(\mathbf{I}_n - \boldsymbol{\beta}\tilde{\mathbf{A}})\hat{\mathbf{x}}(kT+T) = e^{\bar{\mathbf{A}}T}(\mathbf{I}_n + \boldsymbol{\alpha}\tilde{\mathbf{A}})\hat{\mathbf{x}}(kT) + e^{\bar{\mathbf{A}}T}\boldsymbol{\alpha}\mathbf{B}\mathbf{u}(kT) + \boldsymbol{\beta}\mathbf{B}\mathbf{u}(kT+T) \tag{3.4.11}$$

$$\hat{\mathbf{y}}(kT) = \mathbf{C}\hat{\mathbf{x}}(kT) \tag{3.4.12}$$

Introduce

$$\mathbf{p}^T \triangleq (\alpha_1, \cdots, \alpha_n, \beta_1, \cdots, \beta_n)$$

a $2n$-dimensional parameter vector. Define the resulting discrete system defined by equations (3.4.11) and (3.4.12) as $\mathbf{G}(z, \mathbf{p})$. The following result gives the state space representation of $\mathbf{G}(z, \mathbf{p})$ in the general case.

Lemma 3.3 *Suppose a strictly proper analog system $\mathcal{G}(s)$ has a state space representation given by $[\mathbf{A}, \mathbf{B}, \mathbf{C}]$. Then $\mathbf{G}(z, \mathbf{p})$ has a state space representation given by $[\mathbf{F_p}, \mathbf{G_p}, \mathbf{H_p}, \mathbf{J_p}]$ where*

$$\begin{aligned}
\mathbf{F_p} &\triangleq (\mathbf{I}_n - \boldsymbol{\beta}\tilde{\mathbf{A}})^{-1} e^{\tilde{\mathbf{A}}T} (\mathbf{I}_n + \boldsymbol{\alpha}\tilde{\mathbf{A}}) \\
\mathbf{G_p} &\triangleq (\mathbf{I}_n - \boldsymbol{\beta}\tilde{\mathbf{A}})^{-1} e^{\tilde{\mathbf{A}}T} \{\boldsymbol{\alpha}\mathbf{B} + (\mathbf{I}_n + \boldsymbol{\alpha}\tilde{\mathbf{A}})(\mathbf{I}_n - \boldsymbol{\beta}\tilde{\mathbf{A}})^{-1}\boldsymbol{\beta}\mathbf{B}\} \\
\mathbf{H_p} &\triangleq \mathbf{C} \\
\mathbf{J_p} &\triangleq \mathbf{C}(\mathbf{I}_n - \boldsymbol{\beta}\tilde{\mathbf{A}})^{-1}\boldsymbol{\beta}\mathbf{B}
\end{aligned} \qquad (3.4.13)$$

and the matrices $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are given by (3.4.4).

Proof: The result follows by the manipulation of (3.4.11) and (3.4.12) and using the fact that

$$z\mathbf{H}(z\mathbf{I}_n - \mathbf{F})^{-1}\mathbf{G} = \mathbf{H}\mathbf{G} + \mathbf{H}(z\mathbf{I}_n - \mathbf{F})^{-1}\mathbf{F}\mathbf{G}$$

∎

3.4.2 Second Order Case

The above ideas can be carried over to second and higher order integral approximation. For the second order Newton-Cotes approximation, the basic idea is to approximate the integral over two sample periods. This gives a modified Simpson's rule of the form

$$\int_{kT}^{kT+2T} \mathbf{g}(\sigma)\, d\sigma \approx \boldsymbol{\alpha}\mathbf{g}(kT) + \boldsymbol{\beta}\mathbf{g}(kT+T) + \boldsymbol{\gamma}\mathbf{g}(kT+2T) \qquad (3.4.14)$$

The discrete system is now defined in terms of $\mathbf{p}$ with

$$\mathbf{p}^T \triangleq (\alpha_1, ..., \alpha_n, \beta_1, \ldots, \beta_n, \gamma_1, ... \gamma_n)$$

a $3n$-dimensional parameter vector. In addition to $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ defined in (3.4.4), define

$$\boldsymbol{\gamma} \triangleq \mathrm{diag}(\gamma_1, \gamma_2, \ldots, \gamma_m) \qquad (3.4.15)$$

Lemma 3.4 *Suppose an analog system $\mathcal{G}(s)$ has a state space representation $[\mathbf{A}, \mathbf{B}, \mathbf{C}]$ Then the discrete system $\mathbf{G}(z, \mathbf{p})$ resulting from a second order approximation of the integral in*

$$\mathbf{x}(kT+2T) = e^{2\bar{\mathbf{A}}T}\mathbf{x}(kT) + \int_{kT}^{kT+2T} e^{\bar{\mathbf{A}}(kT+2T-\sigma)}\{\tilde{\mathbf{A}}\mathbf{x}(\sigma) + \mathbf{B}\mathbf{u}(\sigma)\}\, d\sigma \tag{3.4.16}$$

is given by

$$\begin{aligned}\hat{\mathbf{x}}(kT+2T) &= \mathbf{A}_1\hat{\mathbf{x}}(kT+T) + \mathbf{A}_0\hat{\mathbf{x}}(kT) + \mathbf{B}_0\mathbf{u}(kT) + \mathbf{B}_1\mathbf{u}(kT+T) \\ &\quad + \mathbf{B}_2\mathbf{u}(kT+2T) &(3.4.17)\\ \hat{\mathbf{y}}(kT) &= \mathbf{C}\hat{\mathbf{x}}(kT) &(3.4.18)\end{aligned}$$

where

$$\begin{aligned}\mathbf{A}_0 &= (\mathbf{I}_n - \gamma\tilde{\mathbf{A}})^{-1}e^{2\bar{\mathbf{A}}T}(\mathbf{I}_n + \boldsymbol{\alpha}\tilde{\mathbf{A}}) &(3.4.19)\\ \mathbf{A}_1 &= (\mathbf{I}_n - \gamma\tilde{\mathbf{A}})^{-1}e^{\bar{\mathbf{A}}T}\boldsymbol{\beta}\tilde{\mathbf{A}} &(3.4.20)\\ \mathbf{B}_0 &= (\mathbf{I}_n - \gamma\tilde{\mathbf{A}})^{-1}e^{2\bar{\mathbf{A}}T}\boldsymbol{\alpha}\mathbf{B} &(3.4.21)\\ \mathbf{B}_1 &= (\mathbf{I}_n - \gamma\tilde{\mathbf{A}})^{-1}e^{\bar{\mathbf{A}}T}\boldsymbol{\beta}\mathbf{B} &(3.4.22)\\ \mathbf{B}_2 &= (\mathbf{I}_n - \gamma\tilde{\mathbf{A}})^{-1}\gamma\mathbf{B} &(3.4.23)\end{aligned}$$

and $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are given by (3.4.4) and (3.4.15). Additionally, define a transformation:

$$\begin{aligned}\overset{(1)}{\mathbf{p}}(kT) &= \hat{\mathbf{x}}(kT) - \mathbf{B}_2\mathbf{u}(kT) &(3.4.24)\\ \overset{(2)}{\mathbf{p}}(kT) &= \hat{\mathbf{x}}(kT+T) - \mathbf{B}_2\mathbf{u}(kT+T) - (\mathbf{B}_1 + \mathbf{A}_1\mathbf{B}_2)\mathbf{u}(kT) &(3.4.25)\end{aligned}$$

and matrices

$$\begin{aligned}\tilde{\mathbf{A}}_1 &= \mathbf{B}_1 + \mathbf{A}_1\mathbf{B}_2 &(3.4.26)\\ \tilde{\mathbf{A}}_2 &= \mathbf{B}_0 + \mathbf{A}_0\mathbf{B}_2 + \mathbf{A}_1\mathbf{B}_1 + \mathbf{A}_1^2\mathbf{B}_2 &(3.4.27)\end{aligned}$$

Then the second order system given by (3.4.17)-(3.4.23) is equivalent to the first order system

$$\begin{bmatrix} \overset{(1)}{\mathbf{p}}(kT+T) \\ \overset{(2)}{\mathbf{p}}(kT+T) \end{bmatrix} = \begin{bmatrix} \mathbf{0} & \mathbf{I}_n \\ \mathbf{A}_0 & \mathbf{A}_1 \end{bmatrix} \begin{bmatrix} \overset{(1)}{\mathbf{p}}(kT) \\ \overset{(2)}{\mathbf{p}}(kT) \end{bmatrix} + \begin{bmatrix} \tilde{\mathbf{A}}_1 \\ \tilde{\mathbf{A}}_2 \end{bmatrix} \mathbf{u}(kT) \tag{3.4.28}$$

$$\mathbf{y}(kT) = \begin{bmatrix} \mathbf{C} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \overset{(1)}{\mathbf{p}}(kT) \\ \overset{(2)}{\mathbf{p}}(kT) \end{bmatrix} + \mathbf{C}\mathbf{B}_2\mathbf{u}(kT) \tag{3.4.29}$$

Proof: The proof closely mimics the first order case. ▮

3.4.3 Mixed Order Case

In the previous sections, all integrals were replaced by either a first or second order method. One can of course envisage mixing first and second order approximations—this can easily be done using the above framework.

Introduce a diagonal matrix $\mathbf{\Lambda}$ which has zeros and ones on the diagonal. Transform the diagonal matrices $\boldsymbol{\beta}$ and γ according to

$$\begin{aligned} \boldsymbol{\beta} &\Rightarrow \boldsymbol{\beta} + (\mathbf{I}_n - \mathbf{\Lambda})\boldsymbol{\alpha} && (3.4.30) \\ \gamma &\Rightarrow \boldsymbol{\beta} + \mathbf{\Lambda}(\boldsymbol{\alpha} + \boldsymbol{\beta}) && (3.4.31) \end{aligned}$$

Under this transformation with special conditions on $\bar{\mathbf{A}}$ (see Section 3.5) and particular state space structures, some of the second order digital integrators resulting from the application of Lemma 3.4 are reduced to first order. In particular, a zero on the i^{th} $(1 \leq i \leq n)$ diagonal element in $\mathbf{\Lambda}$ will result in the i^{th} digital integrator being first order, and a one on the j^{th} $(1 \leq j \leq n)$ diagonal element in $\mathbf{\Lambda}$ will result in the j^{th} digital integrator being second order.

3.4.4 The Open-Loop Problem Statement

The open-loop problem statement of Chapter 2 (c.f. equation (2.2.3)) is now restated in view of the preceeding theory. The optimization problem is one of finding the parameter $\mathbf{p}$ which achieves the minimization:

$$\min_{\mathbf{p}} \mathcal{J}_N(\mathbf{p})$$

where

$$\mathcal{J}_N(\mathbf{p}) = \sum_{k=0}^{N} \|\mathbf{y}(kT) - \hat{\mathbf{y}}_p(kT)\|_2^2 \qquad (3.4.32)$$

The optimization may be performed over the different orders of discretization developed in Sections 3.4.1–3.4.3. For a given $\mathbf{p}$, there corresponds a discrete system $\mathbf{G}(z, \mathbf{p})$. The optimal discrete system is labelled $\mathbf{G}(z, \mathbf{p}^*)$.

3.5 Special Cases to Gain Insight into the Problem

This section deals with the selection of $\bar{\mathbf{A}}$. For particular choices of $\bar{\mathbf{A}}$, special properties manifest themselves. Of particular interest is: whether or not the cost function is convex in $\mathbf{p}$; whether stability of the analog system is preserved under discretization; whether an elegant discretization error analysis exists, in which case an evaluation of performance is possible; and whether mixed order integral approximation is possible. These ideas are expanded upon in what follows. For an arbitrary choice of $\bar{\mathbf{A}}$, none of the listed properties are guaranteed. Only for special selections of $\bar{\mathbf{A}}$ do some of these properties become apparent.

3.5.1 Special Case $\bar{\mathbf{A}} = \mathbf{0}$

In the first order case, equation (3.4.8) reduces to

$$\mathbf{x}(kT+T) = \mathbf{x}(kT) + \int_{kT}^{kT+T} A\mathbf{x}(\sigma) + \mathbf{B}\mathbf{u}(\sigma)\, d\sigma . \tag{3.5.1}$$

In fact, equations (3.4.5) and (3.4.6) can be written as

$$\begin{aligned} \dot{x}(t) &= v_k(t)\,;\ \ k = 1, 2, \cdots, n & (3.5.2)\\ v_k(t) &= \sum_{j=1}^{n} a_{kj}x_j(t) + \sum_{i=1}^{p} b_{ki}u_i(t) \quad (p = \text{number of inputs}) & (3.5.3)\\ y_k(t) &= \sum_{j=1}^{n} c_{kj}x_j(t) & (3.5.4) \end{aligned}$$

Then a digital system $\mathbf{G}(z, \mathbf{p})$ generated via the approximation to the integral in (3.5.1) can be defined by the state space representation

$$\begin{aligned} \hat{x}_m(kT+T) - \hat{x}_m(kT) &= \boldsymbol{\alpha}_m\hat{v}_m(kT) + \boldsymbol{\beta}_m\hat{v}_m(kT+T) & (3.5.5)\\ \hat{v}_m(kT) &= \sum_{j=1}^{n} a_{mj}\hat{x}_j(kT) + \sum_{i=1}^{p} b_{mi}\hat{u}_i(kT) & (3.5.6)\\ \hat{y}_m(kT) &= \sum_{j=1}^{n} c_{mj}\hat{x}_j(kT) & (3.5.7) \end{aligned}$$

In this case, the discretization scheme corresponds to some of the classical methods of discretization. Recall that under the bilinear transformation, each integrator s^{-1} in the continuous-time system is replaced by $0.5T(z+1)(z-1)^{-1}$ to give an approximate discrete-time system. When $\bar{\mathbf{A}} = \mathbf{0}$, each integrator of the n-dimensional system is replaced by

$$\frac{\boldsymbol{\beta}_i z + \boldsymbol{\alpha}_i}{z-1} \quad i = 1, 2, ..., n$$

In particular, when

$$\boldsymbol{\alpha} = \boldsymbol{\beta} = \frac{T}{2}\mathbf{I}_n$$

$\mathbf{G}(z,\mathbf{p})$ is a *bilinear transformation* of $\mathcal{G}(s)$, when

$$\boldsymbol{\alpha} = T\mathbf{I}_n; \quad \boldsymbol{\beta} = 0\mathbf{I}_n$$

$\mathbf{G}(z,\mathbf{p})$ is a *forward difference approximation* of $\mathcal{G}(s)$, and when

$$\boldsymbol{\alpha} = 0\mathbf{I}_n; \quad \boldsymbol{\beta} = T\mathbf{I}_n$$

$\mathbf{G}(z,\mathbf{p})$ is a *backward transformation* of $\mathcal{G}(s)$.

It is well known that the bilinear transformation maps a stable continuous-time system into a stable discrete-time system. However for arbitrary $\mathbf{p}$ (i.e. an arbitrary choice of α_i, β_i in the first order case), a stable $\boldsymbol{\mathcal{G}}(s)$ does not guarantee a stable $\mathbf{G}(z,\mathbf{p})$. Consider

$$\mathbf{G}(z,\mathbf{p}) = \boldsymbol{\mathcal{G}}(f^{-1}(z)) \tag{3.5.8}$$

for some invertible mapping $f(\cdot)$ of the s-domain into the z-domain. It is of interest to know the conditions on $f(\cdot)$ for the preservation of stability of $\boldsymbol{\mathcal{G}}(s)$. The following results consider this problem for particular $f(\cdot)$.

Lemma 3.5 *Let a continuous-time system $\boldsymbol{\mathcal{G}}(s)$ be mapped into a discrete-time system according to equation (3.5.8) where*

$$f(z) = \frac{(T-\alpha)z+\alpha}{z-1}, \quad \alpha \in \mathbb{R}$$

Then

$$f(\{s : \mathbb{R}(s) < 0\}) \subset \{z : |z| < 1\}$$

if and only if $\alpha \leq T/2$.

Proof: This proof and the next use the Schur-Cohn Theorem found in [57]. Without loss of generality assume $T = 1$. The mapping is generated according to

$$\frac{1}{s} = \frac{(1-\alpha)z+\alpha}{z-1} \tag{3.5.9}$$

which can be rewritten as

$$(\frac{1}{s}+\alpha-1)z - \frac{1}{s} - \alpha = 0 \tag{3.5.10}$$

Let $s = \sigma + j\omega$ with $\sigma < 0$. For $|z| < 1$, $\forall\sigma < 0, \omega$ one must have

$$\begin{aligned} \begin{vmatrix} -\alpha - \frac{1}{\sigma+j\omega} & \frac{1}{\sigma+j\omega} + \alpha - 1 \\ \frac{1}{\sigma-j\omega} + \alpha - 1 & -\alpha - \frac{1}{\sigma-j\omega} \end{vmatrix} &= \frac{2\alpha\omega^2 + 2\alpha\sigma^2 + 2\sigma - \sigma^2 - \omega^2}{\sigma^2 + \omega^2} && (3.5.11) \\ &< 0 && (3.5.12) \end{aligned}$$

That is

$$\alpha < \frac{1}{2}(1 - \frac{2\sigma}{\sigma^2 + \omega^2}) \tag{3.5.13}$$

i.e. if and only if $\alpha \leq 1/2$. ▮

The bilinear transformation and the backward Euler rule satisfy the conditions of Lemma 3.5. The following result is for the second order case.

Lemma 3.6 *Let a continuous-time system $\mathcal{G}(s)$ be mapped into a discrete-time system according to equation (3.5.8) where*

$$f(z) = \frac{(2T - \alpha - \beta)z^2 + \beta z + \alpha}{z^2 - 1}, \quad \alpha, \beta \in \mathbb{R}$$

Then

$$f(\{s : \Re(s) < 0\}) \subset \{z : |z| < 1\}$$

if and only if α, β satisfy

$$\{\alpha, \beta \ : \ \{2\alpha + \beta \leq 2\} \cap \{\beta \leq 1\}\} \tag{3.5.14}$$

which corresponds to the shaded area of Figure 3-5.

Proof: The proof proceeds similar to the first order case. From the Schur-Cohn Theorem, the roots of the equation

$$\left(\frac{1}{s} + \alpha + \beta - 2\right) z^2 - \beta z - \frac{1}{s} - \alpha = 0 \tag{3.5.15}$$

lie inside the unit circle $\forall s = \sigma + j\omega$, $\sigma < 0$ if and only if

$$\begin{vmatrix} -\alpha - \frac{1}{\sigma+j\omega} & \frac{1}{\sigma+j\omega} + \alpha + \beta - 2 \\ \frac{1}{\sigma-j\omega} + \alpha + \beta - 2 & -\alpha - \frac{1}{\sigma-j\omega} \end{vmatrix} < 0 \tag{3.5.16}$$

and

$$\begin{vmatrix} -\frac{1}{\sigma+j\omega} - \alpha & 0 & \alpha + \beta - 2 + \frac{1}{\sigma+j\omega} & -\beta \\ -\beta & -\frac{1}{\sigma+j\omega} - \alpha & 0 & \alpha + \beta - 2 + \frac{1}{\sigma+j\omega} \\ \alpha + \beta - 2 + \frac{1}{\sigma-j\omega} & 0 & -\frac{1}{\sigma-j\omega} - \alpha & -\beta \\ -\beta & \alpha + \beta - 2 + \frac{1}{\sigma-j\omega} & 0 & -\frac{1}{\sigma-j\omega} - \alpha \end{vmatrix} > 0$$

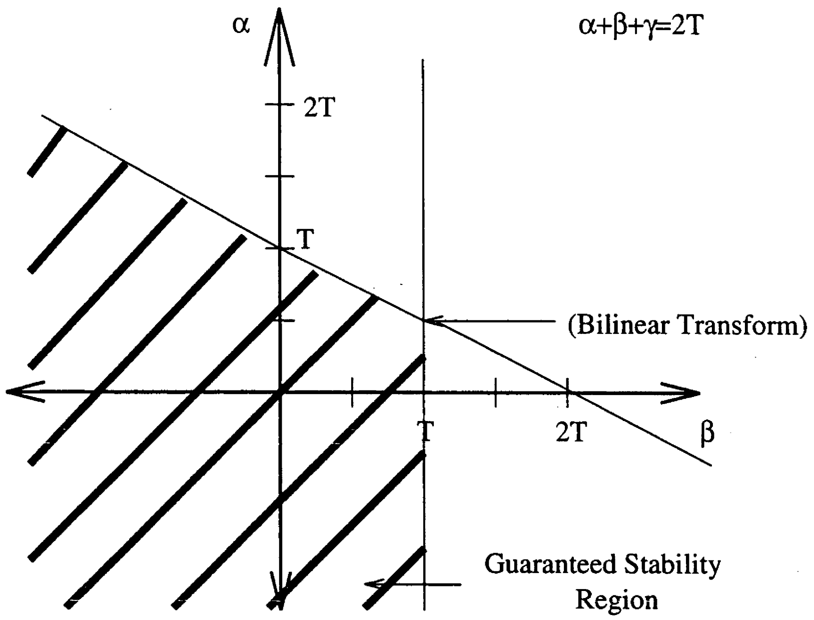

Figure 3-5: Stability region when $\bar{\mathbf{A}} = \mathbf{0}$

(3.5.17)

The first inequality is satisfied in the region described by

$$\{\alpha, \beta \ : \ \{2\alpha + \beta \leq 2\} \cap \{\beta \leq 2\}\} \cup \{\{2\alpha + \beta \geq 2\} \cap \{\beta \geq 2\}\} \tag{3.5.18}$$

and the second inequality by

$$\{\alpha, \beta \ : \ \beta \leq 1\} \tag{3.5.19}$$

which correspond to the region described by the set (3.5.14). ▮

It is interesting to note that Simpson's rule with $\alpha = T/3, \beta = 4T/3, \gamma = T/3$ does not satisfy the conditions of the lemma. The special case $\bar{\mathbf{A}} = \mathbf{0}$ allows a formulation in which the integral associated with each state may be approximated with any order discrete approximation. This special case also has the advantage in effectively being an extension of the traditional methods of discretization. A further advantage of the case $\bar{\mathbf{A}} = \mathbf{0}$ is that the integrators of $\mathcal{G}(s)$ are preserved in $\mathbf{G}(\mathbf{p}, z)$. This fact is a consequence of the particular structure of the transformation; that is, the presence of $z - 1$ in the denominator. Signal invariant transformations and other discretization schemes do not necessarily have this property. This has important consequences in digital controller re-design.

Simulation studies have shown that the choice $\bar{\mathbf{A}} = \mathbf{0}$ can produce numerical problems due to the fact that $\mathbf{F}_p$ can be singular for some particular choice of the parameter $\mathbf{p}$.

Hankel Singular Values

It was noted in [34] that the Hankel singular values of a continuous-time system are identical to those of the corresponding discrete-time system produced by the bilinear transformation. In this section the relationship between the Hankel singular values and the parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ is explored. To enable the analysis, the first order case is considered with $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ restricted to be scalars.

For the case $\bar{\mathbf{A}} = \mathbf{0}$, the discrete system is given by $\mathbf{G}(z, \mathbf{p}) = \mathbf{H}(z\mathbf{I}_n - \mathbf{F})^{-1}\mathbf{G} + \mathbf{J}$ with

$$\begin{aligned}
\mathbf{F} &= (\mathbf{I}_n + \alpha\mathbf{A})(\mathbf{I}_n - \beta\mathbf{A})^{-1} && (3.5.20)\\
\mathbf{G} &= (\alpha\mathbf{I}_n + (\mathbf{I}_n + \alpha\mathbf{A})(\mathbf{I}_n - \beta\mathbf{A})^{-1}\beta)\mathbf{B} && (3.5.21)\\
\mathbf{H} &= \mathbf{C}(\mathbf{I}_n - \beta\mathbf{A})^{-1} && (3.5.22)\\
\mathbf{J} &= \mathbf{C}(\mathbf{I}_n - \beta\mathbf{A})^{-1}\beta\mathbf{B} && (3.5.23)
\end{aligned}$$

The controllability gramian is given by the solution of the discrete Lyapunov equation

$$\mathbf{K} = \mathbf{F}\mathbf{K}\mathbf{F}^T + \mathbf{G}\mathbf{G}^T \tag{3.5.24}$$

Substituting (3.5.20) and (3.5.21) into (3.5.24) and simplifying yields the equation

$$(\alpha - \beta)\mathbf{A}\mathbf{K}\mathbf{A}^T + \mathbf{A}\mathbf{K} + \mathbf{K}\mathbf{A}^T + (\alpha + \beta)\mathbf{B}\mathbf{B}^T = \mathbf{0} \tag{3.5.25}$$

Similarly, substitutions of (3.5.20) and (3.5.22) into the discrete observability Lyapunov equation

$$\mathbf{W} = \mathbf{F}\mathbf{W}\mathbf{F}^T + \mathbf{H}^T\mathbf{H} \tag{3.5.26}$$

gives

$$(\alpha - \beta)\mathbf{A}^T\mathbf{W}\mathbf{A} + \mathbf{A}^T\mathbf{W} + \mathbf{W}\mathbf{A} + \frac{1}{\alpha + \beta}\mathbf{C}^T\mathbf{C} = \mathbf{0} \tag{3.5.27}$$

Lemma 3.7 *Assume $\alpha + \beta = c$ for some constant $c > 0$, and assume $\gamma' \in [0, \alpha - \beta]$. If $|\gamma'\lambda_i(\mathbf{A}) + 1| - 1$ is of one sign for all $i = 1, \ldots, n$, then the solution $\mathbf{K}$ to equation (3.5.25) and $\mathbf{W}$ to equation (3.5.27) satisfy*

$$\mathbf{K} > \mathbf{K}_0 \text{ and } \mathbf{W} > \mathbf{W}_0 \text{ if } |\alpha| > |\beta| \tag{3.5.28}$$

$$\mathbf{K} < \mathbf{K}_0 \text{ and } \mathbf{W} < \mathbf{W}_0 \text{ if } |\alpha| < |\beta| \tag{3.5.29}$$

where $\mathbf{K}_0$ *and* $\mathbf{W}_0$ *are the solutions to equations*

$$\mathbf{A}\mathbf{K}_0 + \mathbf{K}_0\mathbf{A}^T + c\mathbf{B}\mathbf{B}^T \;=\; \mathbf{0}\mathbf{A}^T\mathbf{W}_0 + \mathbf{W}_0\mathbf{A} + \frac{1}{c}\mathbf{C}^T\mathbf{C} = \mathbf{0} \tag{3.5.30}$$

Proof: The result follows from Lemma E.3 in Appendix E with $\gamma = \alpha - \beta$; $\kappa = \alpha + \beta$; and $\kappa = \frac{1}{\alpha+\beta}$. ▮

Theorem 3.2 *Assume* $\gamma' \in [0, \alpha - \beta]$. *If* $|\gamma'\lambda_i(\mathbf{A}) + 1| - 1$ *is of one sign for all* $i = 1, \ldots, n$, *then the solution* $\mathbf{K}$ *to equation (3.5.25) and* $\mathbf{W}$ *to equation (3.5.27) are such that*

$$\mathbf{K}\mathbf{W} > \mathbf{K}_0\mathbf{W}_0 \textit{ and } \mathbf{W} > \mathbf{W}_0 \textit{ if } |\alpha| > |\beta| \tag{3.5.31}$$

$$\mathbf{K}\mathbf{W} < \mathbf{K}_0\mathbf{W}_0 \textit{ and } \mathbf{W} < \mathbf{W}_0 \textit{ if } |\alpha| < |\beta| \tag{3.5.32}$$

where $\mathbf{K}_0$ *and* $\mathbf{W}_0$ *are the solutions to equations 3.5.30 and 3.5.30 respectively.*

Proof: Note that equations (3.5.25) and (3.5.27) are linear equations and so the product $\mathbf{KW}$ is invariant along $\alpha - \beta$ =constant with $\alpha + \beta$ varied. Now any change in α and β can be projected as a change in $\alpha + \beta$ and a change in $\alpha - \beta$. Hence the function $\mathbf{KW}$ is dependent only on the change in $\alpha - \beta$ and the result follows from the previous lemma. ▮

From the preceeding results one can see that starting from the bilinear transformation, if α is increased so that the discretization tends to the forward Euler scheme, then both the controllability and observability gramians will both increase. As a result

$$\text{trace}(\mathbf{KW}) > \text{trace}(\mathbf{K}_0\mathbf{W}_0)$$

That is, the sum of the singular valued squared is increased. Conversely, as α is decreased the sum of the singular valued squared is decreased.

3.5.2 Special Case $\bar{\mathbf{A}} = \mathbf{A}$

In the case $\bar{\mathbf{A}} = \mathbf{A}$, equation (3.4.8) becomes

$$\mathbf{x}(kT+T) = e^{\mathbf{A}T}\mathbf{x}(kT) + \int_{kT}^{kT+T} e^{\mathbf{A}(kT+T-s)}\mathbf{B}\mathbf{u}(s)\,ds \tag{3.5.33}$$

and Lemma 3.3 is greatly simplified with

$$\mathbf{F}_p = e^{\mathbf{A}T},\ \mathbf{G}_p = e^{\mathbf{A}T}(\boldsymbol{\alpha}+\boldsymbol{\beta})\mathbf{B},\ \mathbf{H}_p = \mathbf{C},\ \mathbf{J}_p = \mathbf{C}\boldsymbol{\beta}\mathbf{B}$$

The following result can be stated about the convexity of the cost function in the case $\bar{\mathbf{A}} = \mathbf{A}$. Only the first order case is stated but higher order results follow naturally.

Lemma 3.8 *Given the continuous-time system*

$$\mathbf{x}(kT+T) = e^{\mathbf{A}T}\mathbf{x}(kT) + \int_{kT}^{kT+T} e^{\mathbf{A}(kT+T-\sigma)}\mathbf{B}\mathbf{u}(\sigma)\,d\sigma; \quad \mathbf{y}(kT) = \mathbf{C}\mathbf{x}(kT)$$

and the discrete system

$$\hat{\mathbf{x}}(kT+T) = e^{\mathbf{A}T}\hat{\mathbf{x}}(kT) + e^{\mathbf{A}T}\boldsymbol{\alpha}\mathbf{B}\mathbf{u}(kT) + \boldsymbol{\beta}\mathbf{B}\mathbf{u}(kT+T); \quad \hat{\mathbf{y}}(kT) = \mathbf{C}\hat{\mathbf{x}}(kT)$$

the cost function $J(\mathbf{p})$ given by

$$\mathcal{J}_N(\mathbf{p}) = \sum_{k=0}^{N} \|\mathbf{y}(kT) - \hat{\mathbf{y}}_p(kT)\|_2^2$$

is convex in $\mathbf{p}$.

Proof: The result follows simply from the fact that $\hat{\mathbf{y}}_p(kT)$ is linear in the parameter $\mathbf{p}$ and so $\mathcal{J}_N(\mathbf{p})$ is quadratic in $\mathbf{p}$. The positivity of $\mathcal{J}_N(\mathbf{p})$ then implies convexity in $\mathbf{p}$. ▮

When $\bar{\mathbf{A}} = \mathbf{A}$, preservation of stability for all parameter values $\mathbf{p}$ is guaranteed. This immediately follows from the fact that the eigenvalues of the analog system are mapped according to $e^{n\lambda_i T}$. Here λ_i denotes the eigenvalues of the analog system, and n denotes the order of the integral approximation.

Two other observations about the case $\bar{\mathbf{A}} = \mathbf{A}$ are made at this point. The first is that this case is amenable to error analysis; these results appear in Section 3.6. Second is the fact that there are restrictions on which states can be approximated with first order and which states with second order approximations. States corresponding to eigenvalues occurring in conjugate pairs must be approximated either by both first order or both second order, but not a mixture of first and second order. This is due to the fact that eigenvalues corresponding to the analog system are mapped according to $e^{n\lambda_i T}$. In the case of a conjugate pair of eigenvalues, the eigenvalues can not be "dissected" by trying to make one correspond to a first order approximation and the other correspond to second order. This is in some ways similar to standard model order reduction via balanced truncations where Hankel singular values with multiplicity greater than one can not be "dissected". In the case $\bar{\mathbf{A}} = \mathbf{A}$, for the allocation of first and second order digital integrators to be meaningful, it is assumed that the state space structure of $\mathcal{G}(s)$ is in modal canonical form. That is, for $\mathcal{G}(s) = \mathbf{C}(s\mathbf{I}_n - \mathbf{A})^{-1}\mathbf{B}$, the $\mathbf{A}$ matrix is a block diagonal matrix made up of 1×1 and 2×2 blocks. The 1×1 blocks correspond to real eigenvalues and and 2×2 blocks have eigenvalues in complex conjugate pairs.

3.5.3 Special Case $\bar{\mathbf{A}} = \{a_{ij}\delta_{ij}\}$

Consider the special case, $\bar{\mathbf{A}} = \{a_{ij}\delta_{ij}\}$, where δ_{ij} is the Kronecker delta function. In this case equation (3.4.8) becomes

$$\begin{aligned}
x_1(kT+T) &= e^{a_{11}T}x_1(kT) \\
+ \int_{kT}^{kT+T} & e^{a_{11}(kT+T-\sigma)}\{a_{12}x_2(\sigma) + \cdots + a_{1n}x_n(\sigma) + \sum_{j=1}^{p} b_{1j}u_j(\sigma)\}\, d\sigma \quad (p = \#\text{ inputs}) \\
x_2(kT+T) &= e^{a_{22}T}x_2(kT) \\
+ \int_{kT}^{kT+T} & e^{a_{22}(kT+T-\sigma)}\{a_{21}x_1(\sigma) + a_{23}x_3(\sigma) + \cdots + a_{2n}x_n(\sigma) + \sum_{j=1}^{p} b_{2j}u_j(\sigma)\}\, d\sigma \\
& \cdots \\
& \cdots \\
x_n(kT+T) &= e^{a_{nn}T}x_n(kT) \\
+ \int_{kT}^{kT+T} & e^{a_{nn}(kT+T-\sigma)}\{a_{n1}x_1(\sigma) + a_{n2}x_2(\sigma) + \cdots + a_{n(n-1)}x_{n-1}(\sigma) + \sum_{j=1}^{p} b_{nj}u_j(\sigma)\}\, d\sigma
\end{aligned}$$

Convexity of the cost function and preservation of stability are not guaranteed, and error analysis are not readily amenable to this case. The main advantage of this method is that the integrator corresponding to each state can readily be approximated by either first or second order.

3.6 The Effects of Approximation Order on Discretization Error

Given the task of discretizing an analog system by approximating each analog integrator by a digital approximation, performance advantages may be expected using higher order integral approximations. Increasing the order of the approximations will not degrade the performance of the discretized system. Moreover, experimental evidence confirms that increasing the order of the approximations generally improves the performance of the discretized system. However, it is undesirable to produce very high order digital approximations due to the increased complexity of the discrete-time system.

The quantitative relationship between approximation order and discretization error is a complex one. A complete quantitative analysis of discretization errors produced by first or second order approximations does not appear to be possible. At best, results can be obtained which are in no way elegant and do little to illuminate our understanding. Bounds do exist for special selections of $\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}$—for example, values corresponding to the Newton-Cotes methods. These bounds have been well documented in the literature and can be found in many undergraduate calculus textbooks; see [51, 72] for example.

The error bounds associated with the Newton-Cotes methods are generated using a collocation polynomial. A collocation polynomial is a polynomial which takes on the same functional values at certain points as the original function. These points are called points of collocation. There exists an error bound between the collocation polynomial and the original function, and this bound is used to generate an integral approximation bound.

Unfortunately this technique can not be applied in the general parameter case. The general parameter case uses approximating polynomials which do not necessarily match the original function at any point. As a result, tight error bounds can not be found. In fact, the discretization error can be made arbitrarily large with arbitrary $\mathbf{p}$. Naturally there are additional problems associated with finding discretization error bounds for multivariable systems.

Faced with these difficulties, there are two options available. The first is to consider some simpler problems and try to draw some conclusions from them. In the rest of this section, a selection of simpler problems are presented. The second approach is to perform extensive simulation work. This has been done and a selection of results are presented in Section 3.9.

Of ultimate interest is the generation of a scheme, which, for a given state space realization, allows the designer to select which states require first order integral approximation and which states warrant second order. The factors which effect discretization error, and hence enable the designer to make a sensible allocation, are:

1. Sampling period
2. Hankel singular values
3. Bandwidth (with related undamped natural frequency and damping ratio)
4. Input spectrum

The relationship between these properties and discretization error are explained in Section 3.6.4. First some analytic results are obtained.

3.6.1 Scalar Result

Define the discretization error at the output by

$$E_{output} \triangleq \sup_{k} |y(kT) - \hat{y}(kT)| \tag{3.6.1}$$

The following result gives a bound on this error.

Lemma 3.9 *Consider the scalar system*

$$\begin{aligned} \dot{x}(t) &= -ax(t) + bu(t), \quad a > 0 && (3.6.2) \\ y(t) &= cx(t) && (3.6.3) \end{aligned}$$

Let the integral be approximated via the trapezoidal rule (i.e. $\bar{\mathbf{A}} = \mathbf{A}$ *with* $\alpha = \beta = \frac{T}{2}$*) and let the applied input be*

$$u(t) = U\cos(\omega t), \;\; \omega \geq 0, \;\; A > 0 \tag{3.6.4}$$

Then

$$E_{output} \leq \frac{1}{6}\sigma U(\omega T + aT)^2 + O(T^3) \tag{3.6.5}$$

where σ is the Hankel singular value.

Proof: See Appendix D.1. ▮

By similar arguments, the discretization error bound at the output in the case when the second order Simpson's rule ($\alpha = \gamma = \frac{T}{3}, \beta = \frac{4T}{3}$) is used as the integral approximation, is given by

$$E_{output} \leq \frac{1}{45}\sigma U(\omega T + aT)^4 + O(T^5) \tag{3.6.6}$$

The error formula in both cases is made up of three components:

1. A multiplicative term σ, the Hankel singular value which is related to the "importance" of a particular state.
2. A term ωT related to the steady state error. Recall that the Nyquist sampling theorem says that the sampling period should satisfy $\omega T < \frac{1}{2}$.
3. A term aT related to the transient response of the system. In fact, a^{-1} is the time constant of the system so a sensible choice of sampling period should have $aT < \frac{1}{2}$.

One sees that provided $(\omega + a)T < 1$, the second order approximation yields better results. Furthermore, the error is proportional to the Hankel singular value.

3.6.2 Second Order Result

The previous analysis is now repeated in the case of a second order system.

Lemma 3.10 *Let $\bar{\mathbf{A}} = \mathbf{A}$ and consider a second order system in the form*

$$\dot{\mathbf{x}}(t) = \mathbf{A}\mathbf{x}(t) + \mathbf{b}u(t) \tag{3.6.7}$$

$$y(t) = \mathbf{c}^T\mathbf{x}(t) \tag{3.6.8}$$

with a real canonical realization

$$A = \begin{bmatrix} -a_1 & a_2 \\ -a_2 & -a_1 \end{bmatrix} ; \mathbf{b} = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} ; \mathbf{c}^T = \begin{bmatrix} 1 & 0 \end{bmatrix}, \ (a_1 > 0) \tag{3.6.9}$$

Let the integrals be approximated via the trapezoidal rule, and let the applied input be

$$u(t) = U\cos(\omega t) \tag{3.6.10}$$

Then

$$E_{output} \leq \frac{1}{3}U\frac{\sqrt{b_1^2+b_2^2}}{2a_1}(\omega T + (a_1 + |a_2|)T)^2 \tag{3.6.11}$$

Proof: See Appendix D.2. ▮

The quantity $0.5a_1^{-1}\sqrt{b_1^2+b_2^2}$ is of some interest. From the earlier discussion of the scalar example, one might expect that it is related to the Hankel singular values. The analytic expression for the Hankel singular values of the system at hand is very unattractive, taking up about 10 lines of Maple[1] output. However in the limit as $a_2 \to \infty$, the sum of the Hankel singular values is in fact equal to $0.5a_1^{-1}\sqrt{b_1^2+b_2^2}$. The discrepancy for other smaller values of a_2 is not great, as seen by the following experiment. A number (10000) of random systems were generated with b_1, b_2, a_2 selected to be normally distributed random variables with zero mean, variance equal to 1 (i.e. a_2 small compared to ∞). Also a_1 was created by taking the absolute value of a normally distributed random variable with zero mean, variance equal to 1. A histogram of the ratio

$$\frac{\sum_{i=1}^{2}\sigma_i}{\sqrt{b_1^2+b_2^2/2a_1}}$$

is shown in Figure 3-6. The figures shows that a high proportion of systems have a ratio approaching one, i.e the quantity $0.5a_1^{-1}\sqrt{b_1^2+b_2^2}$ is related to the sum of the Hankel singular values.

A second point of interest is the appearance of the quantity $a_1 + |a_2|$ which is related to the magnitude of the eigenvalues of $\mathbf{A}$, or the undamped natural frequency of the system.

Finally, consider the canonical second order system with transfer function

$$\frac{\omega_n^2}{s^2 + 2\zeta\omega_n s + \omega_n^2}$$

[1] Maple is a registered trademark of Waterloo Maple Software.

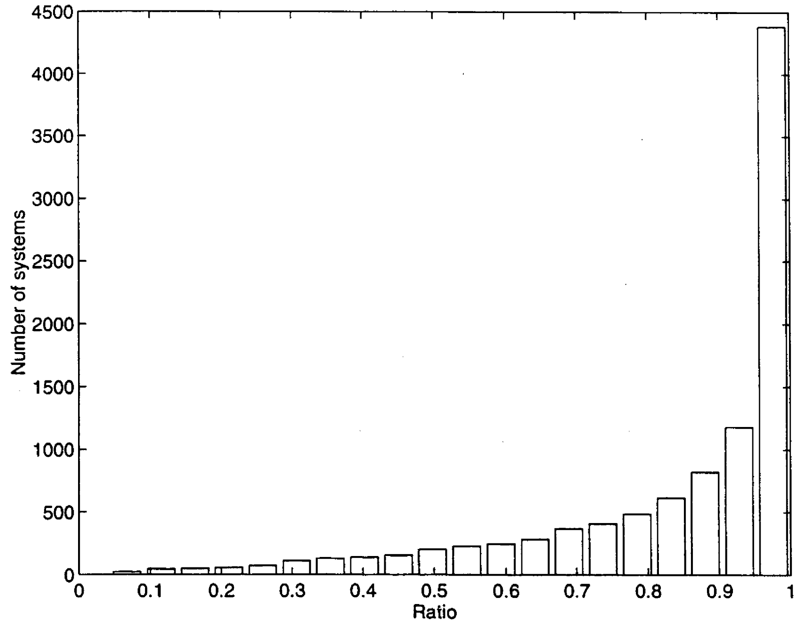

Figure 3-6: Relationship of the error bound to the sum of Hankel singular values

where ω_n is the undamped natural frequency and ζ is the damping ration. One state space realization corresponds to the choices

$$a_1 = \zeta\omega_n,\ a_2 = \omega_n\sqrt{1-\zeta^2},\ b_1 = 0,\ b_2 = \frac{\omega_n}{\sqrt{1-\zeta^2}}$$

in the system described in (3.6.9). In this case (3.6.11) becomes

$$E_{output} \leq \frac{1}{12}U\frac{1}{2\zeta\sqrt{1-\zeta^2}}(\omega T + \omega_n(\zeta + \sqrt{1-\zeta^2})T)^2 \tag{3.6.12}$$

Notice that (3.6.12) is unbounded in the undamped ($\zeta = 0$) and critically damped ($\zeta = 1$) cases and unmeaningful for the overdamped ($\zeta > 1$) case. The unboundedness for the $\zeta = 1$ case is an artefact of the state space realization. Specifically, since there are no oscillatory modes when $\zeta = 1$, the realization is inappropriate and unboundedness occurs. This is an indication of the importance of a good state space realization!

The $\zeta = 0$ result is to be expected. As $\zeta \to 0$ the system approaches instability and the errors become unbounded as signals get out of phase. This is an indication that higher order discretization methods may be appropriate for systems with low damping and corresponding large signals.

3.6.3 Comparing Discrete Systems Approach

There is an inherent difficulty in comparing the difference between a discrete-time system and a continuous-time one. In this section, the error between a discrete-time system and a fast sampled discrete-time system is examined. Suppose a SISO system $\mathcal{G}(s)$ has a state space representation given by

$$\begin{aligned} \dot{\mathbf{x}}(t) &= \mathbf{A}\mathbf{x}(t) + \mathbf{b}u(t) \quad \lambda_i(\mathbf{A}) < 0 && (3.6.13) \\ y(t) &= \mathbf{c}^T\mathbf{x}(t) && (3.6.14) \end{aligned}$$

Denote a system discretized according to Lemma 3.3, with sampling period T, by

$$G(z, \mathbf{p}, T)$$

Denote an error system by $G_e(z, \mathbf{p}, N)$, corresponding to the difference between a discrete-time system at a sampling rate T, and a discrete-time system at a sampling rate $\frac{T}{N}$; $N > 0$, i.e.

$$G_e(z, \mathbf{p}, N) \triangleq G(z, \mathbf{p}, T) - G(z, \mathbf{p}, \frac{T}{N}) \qquad (3.6.15)$$

Lemma 3.11 *Let a discrete-time error system $G_e(z, \mathbf{p}, N)$ be formed according to (3.6.15), with $\bar{\mathbf{A}}$ an arbitrary stable matrix, and the constraint*

$$\boldsymbol{\alpha} + \boldsymbol{\beta} = T\mathbf{I}_n \qquad (3.6.16)$$

where T is the sampling rate. Assume $u(t)$ is a stochastic process with

$$E\{u(t)u(t-\tau)\} = e^{-\varsigma\tau}, \quad \varsigma > 0$$

where E denotes expectation.

Then a bound on the H_2 norm of the error system $G_e(z, \mathbf{p}, N)$ as $N \to \infty$, is given by

$$||G_e(z, \mathbf{p}, N)||_2 \leq T^{\frac{1}{2}}\{1 - e^{-\varsigma T}\}||\mathcal{G}(s)||_2 \qquad (3.6.17)$$

Proof: The proof uses "blocking" techniques and is found in Appendix D.3. ∎

Assumption (3.6.16) is not unreasonable, as simulation studies reveal that for T small compared to the system time constant, the optimal parameters do approximate this equation. Furthermore, the bound supplied by Lemma 3.11 is found to be reasonably tight in most examples.

Note that $\varsigma \to \infty$ corresponds to an uncorrelated input signal. In this case, two points can be noted:

- From equation (D.3.13) in Appendix D.3, $||G_e(z, \mathbf{p}, N)||_2$ is in fact the H_2 norm of the step invariant transformation of the original continuous-time system.
- $||G_e(z, \mathbf{p}, N)||_2 \leq T^{\frac{1}{2}}||\mathcal{G}(s)||_2$

Taking the inequality (3.6.17) further yields

$$\begin{aligned}
||G_e(z, \mathbf{p}, N)||_2 &\leq T^{\frac{1}{2}}\{1 - e^{-\varsigma T}\}||\mathcal{G}(s)||_2 \\
&\leq T^{\frac{1}{2}}\{1 - e^{-\varsigma T}\}\sqrt{\frac{\eta(\pi + 4)}{4\pi}}||\mathcal{G}(s)||_\infty \\
&\leq T^{\frac{1}{2}}\{1 - e^{-\varsigma T}\}\sqrt{\frac{\eta(\pi + 4)}{\pi}}\sum_{i=1}^{n}\sigma_i \\
&\leq \min\{\varsigma T^{\frac{3}{2}}, T^{\frac{1}{2}}\}\sqrt{\frac{\eta(\pi + 4)}{\pi}}\sum_{i=1}^{n}\sigma_i \qquad (3.6.18)
\end{aligned}$$

where σ_i are the Hankel singular values of the continuous-time system, and η is a quantity related to the bandwidth given by

$$\eta = \inf\{\eta_1 \ : \ |\mathcal{G}(j\omega)| \leq \left|\frac{\eta_1}{j\omega + \eta_1}\right| ||\mathcal{G}(j\omega)||_\infty \ \forall\, \omega > \eta_1\} \qquad (3.6.19)$$

The discretization error measured in a two norm sense is bounded by a product involving the sampling period, the bandwidth, the sum of Hankel singular values and a factor which relates to the correlation of the input signal.

If the preceeding analysis is extended to the second order discretization, it can be shown that the same bounds are achieved. This is not surprising for two reasons. The bounds calculated are worst case bounds—one would not expect an improvement in going to second order. Also simulation results show that unless the parameters are optimized, second order discretization performs no better than first order.

It is interesting to apply inequality (3.6.17) to a scalar system (with state space realization $[a, b, c]$), in the case when the input signal is a sinusoid. For $u(t) = U\cos(\omega t)$, the autocorrelation is

$$R(\tau) = \frac{1}{2\pi}\int_0^{2\pi} U\cos(\omega t).U\cos(\omega(t - \tau))\ d\tau = \frac{U^2}{2}\cos(\omega\tau)$$

Normalising so that $R(0) = 1$ gives the normalised autocorrelation between sample points equal to $R(T) = \cos(\omega T)$. Making the approximation

$$e^{-\varsigma T} = \lambda^{N-1} \approx R(T) = \cos(\omega T)$$

and using inequality (3.6.17) gives

$$\begin{aligned}
\|G_e(z,\mathbf{p},N)\|_2 &\leq T^{\frac{1}{2}}\{1-\cos(\omega T)\}\|\mathcal{G}(s)\|_2 \\
&\leq \frac{1}{2}T^{\frac{1}{2}}(\omega T)^2\frac{cb}{\sqrt{2a}} \qquad (3.6.20)\\
&= \frac{1}{\sqrt{2}}\sqrt{aT}\,\sigma\,(\omega T)^2 \qquad (3.6.21)
\end{aligned}$$

Interestingly, the quantities aT, σ and ωT appear again.

3.6.4 Summary of Factors which Affect Discretization Error

In Section 3.6 it was stated that there are four predominant factors which appear to affect discretization error. The preceeding analysis and heuristic arguments are now used to explain these factors. In addition, the problem of allocating different order integral approximations is considered and a set of criterion for selecting integral approximation order is outlined.

In the last three analytic results, it can be seen that the discretization error increases with increasing sampling period. The rate at which this increase takes place is dependent upon many factors, including the selection of the error norm used. This is a fairly intuitive result, suggesting that greater gains can be obtained with higher order approximation as T increases, especially when the sampling frequency is slow compared to the cutoff frequency of the system. However, if the sampling frequency is less than the Nyquist sampling frequency, then the improvements with higher order are more dependent upon the particular example.

The discretization error is dependent upon the Hankel singular values. In a sense this is not surprising. It is well known in model order reduction (see [34] for example) that states which are "not important", as measured by a small relative size of the Hankel singular value, may be removed with less penalty (in a system norm sense) as compared with states with large Hankel singular values. The importance of the Hankel singular values is seen in the preceeding results. This suggests that states associated with small Hankel singular values should have their corresponding integrals approximated by first order approximations, whereas the "more important" states, corresponding to large Hankel singular values, benefit proportionally more from higher order approximations.

In Section 3.3, a comparison is made between the bilinear transformation and a particular first order discretization which is optimized with respect to the cost function

(3.4.32). Figures 3-3 and 3-4 show this comparison and demonstrate that as a increases, greater improvements are obtained using the optimized approach. In this motivational example, it should be pointed out that as a is varied the Hankel singular value remains constant. Note also the presence of the term aT in the relationship (3.6.5). These facts along with results (3.6.11) and (3.6.19) suggest that the bandwidth or the predominant undamped natural frequency of the system plays a role in discretization error. Furthermore, result (3.6.12) suggests that larger errors may be expected in the case where $\zeta \to 0$, and so the discretization may benefit from higher order approximations.

The reason that larger bandwidth systems are associated with larger errors is a little counter-intuitive. The reasoning is similar to that of the $\zeta \to 0$ case just mentioned. In general, for a classical low-pass system, the higher the bandwidth the lower the attenuation for an input signal of a given frequency content. For example in the Bode plot of Figure 3-7, system 1 has a gain of around -30 db for an input signal of frequency $\omega = \omega_i$. System 2, which has a larger bandwidth, has a gain of around -2 db at the same input frequency. All factors considered equal, large bandwidth systems have larger internal signals than small bandwidth systems. The larger signals will result in proportionally larger discretization errors. Conversely, smaller errors are associated with input signals having a high frequency content, due to greater attenuation of these signals.

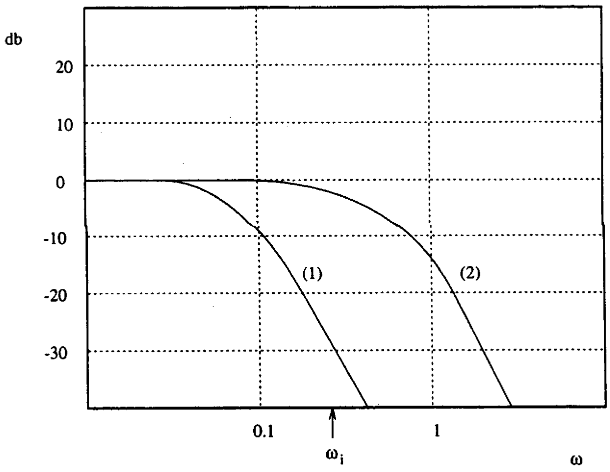

Figure 3-7: Attenuation of signal with $\omega = \omega_i$

3.6.5 Selection Criteria for the Allocation of Approximation Order

A method for allocating different order approximations to different state integrals is presented below. This method is founded upon the factors that affect discretization error that have been outlined in the previous section. These factors have been determined by considering the system as a whole. However, in order to apply this theory, certain sub-systems of a given system are isolated and compared in order to allocate different order approximations. It is assumed that the designer desires to use predominantly first order approximations, but wishes to determine which states justify second order.

The following method has been found to be successful:

1. If $\omega T \ll \frac{1}{2}$ (ω the input frequency) and $\omega_c T \ll \frac{1}{2}$ (ω_c the system's cutoff frequency), significant improvement with higher order approximation is unlikely, so approximate all terms to first order. As a rule of thumb, higher order discretization is has little effect if ωT and $\omega_c T$ are less than 0.05.

2. Determine the Hankel singular values, σ_i. The order of the system can be decreased at this point using standard model order reduction.

3. Next isolate the sub-systems. Given the state matrix $\mathbf{A}$ of the analog system, form a new state space system $[\mathbf{A}, \boldsymbol{\pi}, \boldsymbol{\pi}^T]$ where $\boldsymbol{\pi}$ is a column vector of zeros apart from a one in the position corresponding to the state of interest. This system may be non-minimal, so it is brought into a minimal form with new state matrix $\mathbf{A}'$. The natural frequencies and damping ratios associated with the given state can easily be determined from $\mathbf{A}'$. The bandwidth of the sub-system can be approximated by the largest natural frequency.

   This process is greatly simplified if the system is in modal canonical form.

4. The bandwidths of each of the sub-systems must be weighed against the spectral properties of the input signal. Those sub-systems with large bandwidths are likely to benefit from higher order approximations, especially if they correspond to states with large Hankel singular values. The damping ratios associated with the given state, give additional aid in the selection.

Simulation results which show the validity of this procedure are presented in Section 3.9.

3.7 The Effects of State Space Structure on Discretization Error

Through simulation studies, it has been found that the particular analog state space structure used for discretization has a significant effect upon the quality of the discretization. Inequality (3.6.12) also suggests a correspondence between state space structure and discretization error. Naturally, one would like to find state space structures which reduce discretization errors.

The approach taken at this point is based on a physical problem: noise in an operational amplifier. The argument presented in this section demonstrates that a parallel exists between noise present in an operational amplifier acting as an integrating circuit, and the "noise" induced by approximating an analog integral by a digital approximation.

3.7.1 An Excursion Into Operational Amplifiers

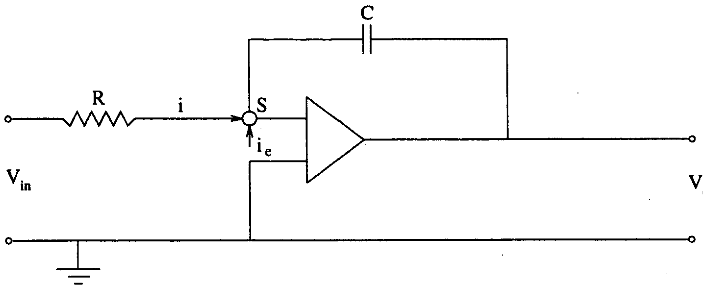

Figure 3-8: Analog integrator with noise

Consider the operational amplifier circuit in Figure 3-8 in which V_{in} is the input voltage, V_{out} is the output voltage, and R and C are the values of resistance and capacitance respectively. Initially, assume $i_e = 0$. Assuming an ideal operational amplifier (infinite input impedance, zero output impedance, high gain), the summing junction S is kept at zero potential (S is a *virtual ground*). The input-output relationship is then given by

$$V_{out} = \frac{1}{C}\int_{t_0}^{t_f} i\, dt \tag{3.7.1}$$

$$= -\frac{1}{RC}\int_{t_0}^{t_f} V_{in}\, dt \tag{3.7.2}$$

and so the output voltage is the integral of the input signal.

One of the properties of a "real" integrating circuit is noise associated with the operational amplifier. This noise is modelled as an independent white, zero mean current source i_e, where the point S is assumed to remain at zero potential. One then has

$$\begin{aligned} V_{out} &= \frac{1}{C}\int_{t_0}^{t_f} i + i_e \, dt && (3.7.3) \\ &= -\frac{1}{RC}\int_{t_0}^{t_f} V_{in} \, dt + \frac{1}{C}\int_{t_0}^{t_f} i_e \, dt && (3.7.4) \end{aligned}$$

For a system

$$\begin{aligned} \dot{x} &= ax + bu && (3.7.5) \\ y &= cx && (3.7.6) \end{aligned}$$

in which the integrator is modelled by the noisy integrating circuit of equation (3.7.4). If the state x is equivalent to V_{out}, then

$$\begin{aligned} \dot{V}_{out} &= -\frac{a}{RC}V_{out} - \frac{b}{RC}u + \frac{1}{C}i_e && (3.7.7) \\ y &= cV_{out} && (3.7.8) \end{aligned}$$

Therefore, using this particular noise model, the integrator noise appears directly in the state equation but does not appear in the output equation.

The minimization of the effects of this integrator noise is a common problem in electronic engineering. For example, a common practical problem is: given n integrators of variable quality, how should one select the position of each integrator as a function of its quality, so as to minimize the effect of integrator noise on the amplifier output?

Proposition 3.1 *The minimization of discretization error is closely associated with the minimization of integrator noise in an analog operational amplifier circuit.*

This claim is supported by the fact that when an digital integrator approximates an analog integrator, an error is introduced. If the input signal into the integrator is "sufficiently rich", then one would expect that the error made from one sampling period to the next would be "reasonably uncorrelated". Heuristically, the discretization process can be thought of as being equivalent to replacing each ideal analog integrator by an analog integrator plus an independent white noise source.

Minimization of discretization error and minimization of integrator noise are not entirely analogous due to the presence of memory associated with integration. The whiteness assumption that will be made on the discretization error does not strictly hold, however simulation studies suggest that the proposition is well supported.

3.7.2 Optimal State Space Structure for the Minimization of Integrator Noise

In this section, the problem of determining the state space structure which minimizes the effect of integrator noise on the output of a given system is studied.

An n-dimensional system is associated with n integrators. It is well known that, for a given input-output transfer function, there are an infinite number of realizations using the n integrators. However if each integrator is not a "pure" integrator but has noise associated with it, then there are some structures which are better than others, in the sense of minimizing the combined effect of the integrator noise at the output.

Assume that the state space model of our system without integrator noise is given by

$$\begin{aligned} \dot{\mathbf{x}}(t) &= \mathbf{A}\mathbf{x}(t) + \mathbf{B}\mathbf{u}(t) && (3.7.9) \\ \mathbf{y}(t) &= \mathbf{C}\mathbf{x}(t) && (3.7.10) \end{aligned}$$

In accord with the Section 3.7.1, associate with each integrator a noise source η_i ($1 \leq i \leq n$). Model the noise as appearing at the input of each integrator with each η_i white, Gaussian and zero mean. Let the variance of noise source η_i be $\omega_i = E\{\eta_i^2\}$, where E denotes expectation. Let

$$\boldsymbol{\eta}^T(t) = \begin{bmatrix} \eta_1(t) & \eta_2(t) & \ldots & \eta_n(t) \end{bmatrix}^T \qquad (3.7.11)$$

with

$$E\{\boldsymbol{\eta}(t)\boldsymbol{\eta}^T(\tau)\} = \mathrm{diag}(\omega_1, \omega_2, \ldots, \omega_n) = \boldsymbol{\Omega}_\eta \; \delta(t-\tau) \qquad (3.7.12)$$

where $\delta(\cdot)$ is the Dirac delta function and $\boldsymbol{\Omega}_\eta$ is a diagonal matrix containing the covariances of each of the $\eta_i(t)$. The state space model which represents the effect of integrator noise is then given by

$$\begin{aligned} \dot{\mathbf{x}}(t) &= \mathbf{A}\mathbf{x}(t) + \boldsymbol{\eta}(t) + \mathbf{B}\mathbf{u}(t) && (3.7.13) \\ \mathbf{y}(t) &= \mathbf{C}\mathbf{x}(t) && (3.7.14) \end{aligned}$$

Let $\mathbf{P}$ denote the the steady state covariance of the state $\mathbf{x}(t)$ due to the integrator noise. The matrix $\mathbf{P}$ satisfies the Lyapunov equation

$$\mathbf{A}\mathbf{P} + \mathbf{P}\mathbf{A}^T + \boldsymbol{\Omega}_\eta = \mathbf{0} \qquad (3.7.15)$$

whereby $\mathbf{P}$ satisfies

$$\mathbf{P} = \int_0^\infty e^{\mathbf{A}t}\boldsymbol{\Omega}_\eta e^{\mathbf{A}^Tt}\ dt \tag{3.7.16}$$

The steady state covariance of the output $\mathbf{y}(t)$ due to $\boldsymbol{\eta}(t)$ is then given by

$$\boldsymbol{\Sigma}_\eta = \mathbf{CPC}^T \tag{3.7.17}$$

Also, assume that the system input $\mathbf{u}(t)$ is an independent white noise process of a given intensity, $\boldsymbol{\Omega}_u$. Assume that the covariance of the output due to $\mathbf{u}(t)$ is given by $\boldsymbol{\Sigma}_u$, and let

$$(\boldsymbol{\Sigma}_u)_{jj} = \nu_j^2\ ;\ \ j = 1, 2, \ldots, n_y \tag{3.7.18}$$

where n_y is the number of outputs. Define

$$\boldsymbol{\mathcal{W}}_u \triangleq \text{diag}(\nu_1^2, \nu_2^2, \ldots, \nu_{n_y}^2) \tag{3.7.19}$$

Finally, define an output noise gain

$$\kappa \triangleq \text{tr}(\boldsymbol{\mathcal{W}}_u^{-1}\boldsymbol{\Sigma}_\eta) \tag{3.7.20}$$

The noise gain reflects the effects of the integrator noise on the output weighted by the external input. The output noise can be calculated from

$$\begin{aligned}
\kappa &= \text{tr}(\boldsymbol{\mathcal{W}}_u^{-1}\mathbf{C}\int_0^\infty e^{\mathbf{A}t}\boldsymbol{\Omega}_\eta e^{\mathbf{A}^Tt}\,dt\ \mathbf{C}^T) && (3.7.21)\\
&= \text{tr}(\boldsymbol{\Omega}_\eta\int_0^\infty e^{\mathbf{A}^Tt}\mathbf{C}^T\boldsymbol{\mathcal{W}}_u^{-1}\mathbf{C}e^{\mathbf{A}t}\,dt) && (3.7.22)\\
&= \text{tr}(\boldsymbol{\Omega}_\eta\int_0^\infty e^{\mathbf{A}^Tt}\tilde{\mathbf{C}}^T\tilde{\mathbf{C}}e^{\mathbf{A}t}\,dt) && (3.7.23)\\
&= \text{tr}(\boldsymbol{\Omega}_\eta\mathbf{Q}) && (3.7.24)
\end{aligned}$$

where $\mathbf{Q}$ satisfies the Lyapunov equation

$$\mathbf{A}^T\mathbf{Q} + \mathbf{QA} + \tilde{\mathbf{C}}^T\tilde{\mathbf{C}} = \mathbf{0} \tag{3.7.25}$$

and $\tilde{\mathbf{C}} \triangleq \mathcal{W}_u^{-1/2}C$].

Let $\mathbf{z}(t) \triangleq \mathbf{T}_z\mathbf{x}(t)$ define a new state space representation of the system given by equations (3.7.13) and (3.7.14). Then it is straightforward to show that the output noise gain κ becomes

$$\kappa = \text{tr}(\boldsymbol{\Omega}_\eta\mathbf{T}_z^T\mathbf{Q}\mathbf{T}_z) \tag{3.7.26}$$

where $\mathbf{Q}$ is given by (3.7.25).

Given $\mathbf{u}(t)$ an independent white noise process of intensity, $\mathbf{\Omega}_u$, a scaling condition is imposed upon the states. The scaling condition is given by

$$\{\ \mathbf{T}_z^{-1}\mathbf{K}(\mathbf{T}_z^{-1})^T\ \}_{jj} = 1\ ;\ \ j = 1, 2, \ldots, n \tag{3.7.27}$$

where

$$\mathbf{AK} + \mathbf{KA}^T + \tilde{\mathbf{B}}\tilde{\mathbf{B}}^T = \mathbf{0}\ ;\ \ \tilde{\mathbf{B}} \overset{\triangle}{=} \mathbf{B}\mathbf{\Omega}_u^{1/2} \tag{3.7.28}$$

Scaling guarantees that the integrator outputs are neither "too small" nor "too large" in the L_2 sense. A small integrator signal means a poor SNR, while a large integrator signal will produce significant overflow and integrator distortion.

The aim of the problem is to find an invertible similarity transformation $\mathbf{T}_z$ such that the output noise gain given by (3.7.26) is minimized according to a scaling condition (3.7.27). This is a well known problem (see for example [84]) whose solution is given by the following lemma.

Lemma 3.12 *Suppose* $[\tilde{\mathbf{A}}(=\mathbf{A}), \tilde{\mathbf{B}}, \tilde{\mathbf{C}}]$ *is a minimal asymptotically stable* n^{th} *order system such that* $[\tilde{\mathbf{A}}, \tilde{\mathbf{B}}, \tilde{\mathbf{C}}]$ *is in input balanced form with*

$$\mathbf{K} = \mathbf{I}_n\ \ ;\ \ \mathbf{Q} = \mathbf{\Sigma}^2 = diag(\sigma_1^2, \sigma_2^2, \ldots, \sigma_n^2).$$

Then the class of state space realizations $[\mathbf{A}, \mathbf{B}, \mathbf{C}]$ *where*

$$\mathbf{A} = \mathbf{T}_z^{-1}\tilde{\mathbf{A}}\mathbf{T}_z;\ \mathbf{B} = \mathbf{T}_z^{-1}\tilde{\mathbf{B}};\ \mathbf{C} = \tilde{\mathbf{C}}\mathbf{T}_z$$

which minimizes (3.7.26) subject to the L_2 *scaling constraint (3.7.27) is defined by the transformation* $\mathbf{T}_z$ *with singular value decomposition*

$$\mathbf{T}_z = \mathbf{U}\mathbf{\Pi}\mathbf{V}^T \tag{3.7.29}$$

in which $\mathbf{U}, \mathbf{\Pi}$ *and* $\mathbf{V}$ *are chosen such that:*

1. $\mathbf{D} = \mathbf{U}^T\mathbf{\Sigma}\mathbf{U}$ *is diagonal*
2. $\mathbf{\Pi}^2 = \bar{\lambda}\mathbf{\Omega}_\eta^{-1}\mathbf{D}^{-1}\ \ ;\ \ \bar{\lambda} = \frac{1}{n}\sum_{k=1}^{n}\sigma_k\omega_k$*; where the* ω_k *are defined by equation (3.7.12)*
3. $\{\mathbf{V}\mathbf{\Omega}_\eta\mathbf{D}\mathbf{V}^T\}_{kk} = \bar{\lambda}\ \ \forall k$

Under these conditions

$$\kappa^* = \arg\min_{\mathbf{T}_z} \kappa = \frac{\bar{\omega}}{n}\left(\sum_{k=1}^{n}\sigma_k\right)^2\ \ ;\ \bar{\omega} = \frac{\sum_{k=1}^{n}\sigma_k\omega_k}{\sum_{k=1}^{n}\sigma_k} \tag{3.7.30}$$

Proof: See [84], result 4.4.5. ▮

In [84], this result is presented in relation to the problem of minimizing the effects of quantization. Furthermore, the result is used to argue the optimal assignment of different wordlengths to each state. It is reasoned that if an assignment of wordlengths is made to mimimize the total output roundoff noise, then the transformation (3.7.29) should be such that $\mathbf{T}_z = \mathbf{I}_n$, i.e. there is strong justification for choosing the input balanced realization. Those states associated with large Hankel singular values should be targeted for reducing the output noise gain by placing a high order digital integrator on the state integral. This again confirms what was stated in the Sections 3.6.4 and 3.6.5.

Simulation results confirm that the state space structure obtained from Lemma 3.12 is effective in reducing discretization error. This is demonstrated in Section 3.9.3, see Table 3.4. Also the input balanced realization is shown to be good.

3.8 The Discretization Algorithm

In this section, the general discretization algorithm which minimizes the cost function $\mathcal{J}_N(\mathbf{p})$ is presented.

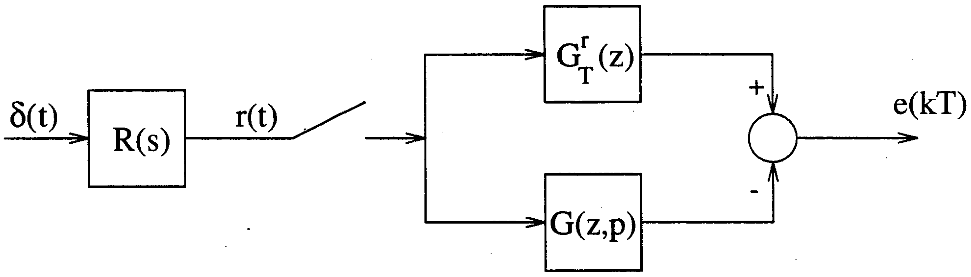

Figure 3-9: Error System

Referring to Figure 3-9, assume the input signal $\mathbf{r}(t)$ is generated via the impulse response of a strictly proper linear time-invariant system $\mathcal{R}(s)$. Denote the signal invariant transformation of $\mathcal{G}(s)$ with respect to $\mathbf{r}(t)$ by $\mathbf{G}_T^r(z)$. Recall the cost function

(3.4.32) written as

$$\mathcal{J}_N(\mathbf{p}) = \sum_{k=0}^{N} \|\mathbf{e}(kT)\|_2^2$$

where $\mathbf{e}(kT)$ is the signal generated by the difference of the responses of $\mathbf{G}(z,\mathbf{p})$ and $\mathbf{G}_T^r(z)$.

Introduce $\mathbf{r}(kT)$, a discrete signal generated by sampling $\mathbf{r}(t)$. The signal $r(kT)$ can be generated via the discrete impulse response of $\mathbf{R}_T(z)$, where $\mathbf{R}_T(z)$ is the impulse invariant transformation of $\mathcal{R}(s)$. In the general second order case, the system $\mathbf{G}(z,\mathbf{p})$ is described by equations (3.4.28) and (3.4.29). Denote a state space realization of the cascade of $\mathbf{G}(z,\mathbf{p})$ and $\mathbf{R}_T(z)$ by $[\mathbf{F}_p,\mathbf{G}_p,\mathbf{H}_p,\mathbf{J}_p]$. Denote a state space realization of the cascade of $\mathbf{G}_T^r(z)$ and $\mathbf{R}_T(z)$ by $[\mathbf{F}_r,\mathbf{G}_r,\mathbf{H}_r,\mathbf{J}_r]$. Form a state space error system $[\mathbf{F},\mathbf{G},\mathbf{H},\mathbf{J}]$ with

$$\begin{aligned} \mathbf{F} &\triangleq \begin{bmatrix} \mathbf{F}_p & \mathbf{0} \\ \mathbf{0} & \mathbf{F}_r \end{bmatrix};\ \mathbf{G} \triangleq \begin{bmatrix} \mathbf{G}_p \\ \mathbf{G}_r \end{bmatrix};\ \mathbf{H} \triangleq [\mathbf{H}_p \ \ -\mathbf{H}_r]; \\ \mathbf{J} &\triangleq \mathbf{J}_p - \mathbf{J}_r \end{aligned} \tag{3.8.1}$$

The signal $\mathbf{e}(kT)$ is generated via the impulse response of the system (3.8.1) i.e.

$$\mathbf{x}(kT+T) = \mathbf{F}\mathbf{x}(kT) + \mathbf{G}\delta(kT) \tag{3.8.2}$$

$$\mathbf{e}(kT) = \mathbf{H}\mathbf{x}(kT) + \mathbf{J}\delta(kT) \tag{3.8.3}$$

where $\delta(\cdot)$ is the Kronecker delta function.

In creating an algorithm for the minimization of $\mathcal{J}_N(\mathbf{p})$, ideally the knowledge of the cost function and its gradient at a point $\mathbf{p}$ should be available. The technique of solving a discrete-time Lyapunov equation can be used if the eigenvalues of $\mathbf{F}$ are guaranteed to lie strictly inside the unit circle. However, this is not always going to be the case, especially if the reference signal model has any poles on the unit circle.

The method found most effective is the doubling algorithm for discrete-time Lyapunov equations. This algorithm is reviewed in [2]. It allows the rapid computation of the quantity

$$\mathbf{X}_{n_{passes}} = \sum_{i=0}^{N} \mathbf{F}^i \mathbf{G}\mathbf{G}^T(\mathbf{F}^T)^i, \ \ N = 2^{n_{passes}} - 1$$

which in turn gives $\mathcal{J}_N(\mathbf{p}) = \mathrm{tr}(\mathbf{H}\mathbf{X}_{n_{passes}}\mathbf{H}^T + \mathbf{J}\mathbf{J}^T)$. The quantity n_{passes} gives the *number of passes* through the algorithm about to be presented. The limit as $N \to \infty$, if it exists, can be calculated very quickly. Furthermore, the partial derivatives of $\mathcal{J}_N(\mathbf{p})$

with respect to the parameters $\mathbf{p}$ can also be calculated quickly using this algorithm. The doubling algorithm with gradient calculations is now outlined.

Algorithm 3.1 (Doubling algorithm)

```
$\mathbf{X}_0 := \mathbf{G}\mathbf{G}^T$
FOR $i = 1$ TO $3n$
    CALCULATE $\frac{\partial \mathbf{X}_0}{\partial \mathbf{p_i}}$
    $\mathbf{W}_{0,i} := \frac{\partial \mathbf{F}}{\partial \mathbf{p}}$
NEXT $i$
$\mathbf{Y}_0 := \mathbf{F}$
FOR $j = 1$ TO $n_{passes}$
    FOR $i = 1$ to $3n$
        $\frac{\partial \mathbf{X}_j}{\partial \mathbf{p_i}} := \mathbf{W}_{j-1}\mathbf{X}_{j-1}\mathbf{Y}^T_{j-1} + \mathbf{Y}_{j-1}\mathbf{X}_{j-1}\mathbf{W}^T_{j-1} + \mathbf{Y}_{j-1}\frac{\partial \mathbf{X}_{j-1}}{\partial \mathbf{p_i}}\mathbf{Y}^T_{j-1} + \frac{\partial \mathbf{X}_{j-1}}{\partial \mathbf{p_i}}$
        $\mathbf{W}_{j,i} := \mathbf{Y}_{j-1}\mathbf{W}_{j-1,i} + \mathbf{W}_{j-1,i}\mathbf{Y}_{j-1}$
    NEXT $i$
    $\mathbf{X}_j := \mathbf{Y}_{j-1}\mathbf{X}_{j-1}\mathbf{Y}^T_{j-1} + \mathbf{X}_{j-1}$
    $\mathbf{Y}_j := \mathbf{Y}^2_{j-1}$
NEXT $j$
```

The cost function is given by

$$\mathcal{J}_N(\mathbf{p}) = \operatorname{tr}(\mathbf{H}\mathbf{X}_{n_{passes}}\mathbf{H}^T + \mathbf{J}\mathbf{J}^T)$$

and the gradient by

$$\frac{\partial \mathcal{J}_N(\mathbf{p})}{\partial \mathbf{p_i}} = \operatorname{tr}\left(\frac{\partial \mathbf{J}}{\partial \mathbf{p_i}}\mathbf{J}^T + \mathbf{J}\frac{\partial \mathbf{J}^T}{\partial \mathbf{p_i}} + \mathbf{H}\frac{\partial \mathbf{X}_{n_{passes}}}{\partial \mathbf{p_i}}\mathbf{H}^T + \frac{\partial \mathbf{H}}{\partial \mathbf{p_i}}\mathbf{X}_{n_{passes}}\mathbf{H}^T + \mathbf{H}\mathbf{X}_{n_{passes}}\frac{\partial \mathbf{H}^T}{\partial \mathbf{p_i}}\right)$$

With this information, the quasi-Newton method with a cubic polynomial line search method available with MATLAB's Optimization Toolbox has been found to be an effective algorithm for the minimization of $\mathcal{J}_N(\mathbf{p})$. This method is fast even for large N and n.

In light of the preceeding results, the following method is proposed.

1. Select $\bar{\mathbf{A}}$—generally $\bar{\mathbf{A}} = \{a_{ij}\delta_{ij}\}$ is a good first choice.

2. Determine a state space realization of the continuous-time system. This can be generated according to Lemma 3.12. Alternatively, an internally balanced

realization is generally quite satisfactory. The selection $\bar{\mathbf{A}} = \mathbf{A}$ with a modal canonical realization has been found to have the best numerical properties for systems of high order.

3. Use the procedure outlined in Section 3.6.5 to determine the importance of each state in relation to discretization—recall that this involves determining the Hankel singular values of the system and the bandwidths of sub-systems. A state may even be removed at this point, as in standard model reduction.

4. Select n_{passes}, the number of passes through the algorithm. Typically $n_{passes} \approx 4-5$ is appropriate if the transient response is the major concern, while $n_{passes} \approx 15$ is suitable if the steady-state performance is important.

5. Initialize the parameter $\mathbf{p}$ to correspond to a standard Newton-Cotes value.

6. Form a state space error system according to equation (3.8.1).

7. Find the optimal $\mathbf{p}$ using the algorithm of this section.

This algorithm has been implemented in MATLAB code; a description can be found in Appendix F.

A further point concerning steady state performance is worth mentioning. Preservation of the frequency characteristics is often the main objective of discrete approximation, and in [15] necessary and sufficient conditions are given for the retention of sinusoidal steady-state properties of the prototype system, as well as conditions for the stability of the discretized system. The algorithm presented here does not have any stability constraints or any mechanism to guarantee the retention of sinusoidal steady-state properties. However in practice this does not seem to be a problem, especially if n_{passes} is large.

3.9 Simulation Results

It should be stressed that many of the statements and conclusions appearing in Sections 3.6 and 3.7 have been drawn from the results of extensive simulation studies. In this section, a selection of these results are presented and it shown that they support the relevant assertions. The results were obtained by using the discretization algorithm of Section 3.8.

3.9.1 Examination of the Effects of Bandwidth upon Discretization Error

In the first study, the relationship between discretization error and bandwidth was investigated. A third order system, comprised of a parallel connection of a canonical second order system with a first order system, was used. The realization $[\mathbf{A}, \mathbf{b}, \mathbf{c}^T]$ used was

$$\mathbf{A} = \begin{bmatrix} 0 & 1 & 0 \\ -w_n^2 & -2\zeta\omega_n & 0 \\ 0 & 0 & -a \end{bmatrix} ; \mathbf{b} = \begin{bmatrix} 0 \\ 1 \\ a \end{bmatrix} ; \mathbf{c}^T = \begin{bmatrix} \omega_n^2 & 0 & c \end{bmatrix}$$

The input was a sum of sinusoids,

$$u(t) = \sin(\omega_1 t) + \cos(\omega_2 t)$$

The optimization was performed over $n_{passes} = 15$ passes, with all initial states equal to zero. Although $\bar{\mathbf{A}} = \mathbf{A}$, the choice of $\bar{\mathbf{A}}$ did not greatly affect the final cost function in this particular study.

A number of experiments were performed, first with $T = 0.08$ seconds. The results are summarized in Tables 3.1.a–3.1.d. The column "HSV" gives the Hankel singular value corresponding to each state, "BW" gives the approximate bandwidth corresponding to each sub-system, and "allocation" gives the order of the integral approximation corresponding to each state.

Trials A1-A4 illustrate the case when the input frequencies are outside the bandwidth of a particular sub-system, specifically the sub-system formed by the first and second states. Notice that increasing the order of discretization has little effect. Trials A5-A8 again illustrate this, however the improvement due to second order discretization is not as great when the product of the input frequencies and the sampling period are significantly less than 0.5; i.e. $\omega_i T \ll 0.5$.

Selecting a higher order discretization on any of the state integrals results in an improvement. This is evidenced in trials A9-A12. Neither sub-system attenuates the input significantly, and $\omega_i T$ is large. Trials A13-A16 are basically the reverse of A1-A4. It is interesting to note that the scalar sub-system with effective damping ratio 1 can still benefit from the use of higher order discretization. This appears to be supported by result (3.6.12).

In Table 3.2, A9-A12 are repeated for a larger sampling period, $T = 0.2$. A large

Trial	ω_1	ω_2	ω_n	ζ	a	c	HSV	BW	Allocation	Error
A1	3	6	0.1	0.7	10	1	0.7056	0.1	1	1.4489
							0.1801	0.1	1	
							0.4745	10	1	
A2	3	6	0.1	0.7	10	1	0.7056	0.1	2	1.4489
							0.1801	0.1	2	
							0.4745	10	1	
A3	3	6	0.1	0.7	10	1	0.7056	0.1	1	0.1073
							0.1801	0.1	1	
							0.4745	10	2	
A4	3	6	0.1	0.7	10	1	0.7056	0.1	2	0.1073
							0.1801	0.1	2	
							0.4745	10	2	

Table 3.1.a: Relationship between discretization error and bandwidth

Trial	ω_1	ω_2	ω_n	ζ	a	c	HSV	BW	Allocation	Error
A5	0.5	1	0.1	0.7	10	1	0.7056	0.1	1	0.1233
							0.1801	0.1	1	
							0.4745	10	1	
A6	0.5	1	0.1	0.7	10	1	0.7056	0.1	2	0.1233
							0.1801	0.1	2	
							0.4745	10	1	
A7	0.5	1	0.1	0.7	10	1	0.7056	0.1	1	0.1044
							0.1801	0.1	1	
							0.4745	10	2	
A8	0.5	1	0.1	0.7	10	1	0.7056	0.1	2	0.1044
							0.1801	0.1	2	
							0.4745	10	2	

Table 3.1.b: Relationship between discretization error and bandwidth

Trial	ω_1	ω_2	ω_n	ζ	a	c	HSV	BW	Allocation	Error
A9	3	6	10	0.7	10	1	1.1480 0.0109 0.1589	10 10 10	1 1 1	1.8392
A10	3	6	10	0.7	10	1	1.1480 0.0109 0.1589	10 10 10	2 2 1	0.3258
A11	3	6	10	0.7	10	1	1.1480 0.0109 0.1589	10 10 10	1 1 2	0.2800
A12	3	6	10	0.7	10	1	1.1480 0.0109 0.1589	10 10 10	2 2 2	0.2740

Table 3.1.c: Relationship between discretization error and bandwidth

Trial	ω_1	ω_2	ω_n	ζ	a	c	HSV	BW	Allocation	Error
A13	3	6	10	0.7	0.1	1	0.7460 0.1844 0.4384	10 10 0.1	1 1 1	0.0802
A14	3	6	10	0.7	0.1	1	0.7460 0.1844 0.4384	10 10 0.1	2 2 1	0.0715
A15	3	6	10	0.7	0.1	1	0.7460 0.1844 0.4384	10 10 0.1	1 1 2	0.0741
A16	3	6	10	0.7	0.1	1	0.7460 0.1844 0.4384	10 10 0.1	2 2 2	0.0715

Table 3.1.d: Relationship between discretization error and bandwidth

Trial	ω_1	ω_2	ω_n	ζ	a	c	HSV	BW	Allocation	Error
B1	3	6	10	0.7	10	1	1.1480 0.0109 0.1589	10 10 10	1 1 1	56.7932
B2	3	6	10	0.7	10	1	1.1480 0.0109 0.1589	10 10 10	2 2 1	0.6844
B3	3	6	10	0.7	10	1	1.1480 0.0109 0.1589	10 10 10	1 1 2	0.6739
B4	3	6	10	0.7	10	1	1.1480 0.0109 0.1589	10 10 10	2 2 2	0.6687

Table 3.2: Relationship between discretization error and bandwidth

improvement is seen in going to higher order approximations.

3.9.2 Examination of the Effects of the Hankel Singular Values upon Discretization Error

Next, the importance of the Hankel singular values was examined. The system of interest was given by

$$\mathcal{G}(s) = \frac{s+2+\epsilon}{(s+2)(s+1)}$$

Realization 1 had a state space structure given by

$$\mathbf{A} = \begin{bmatrix} -1 & 0 \\ \epsilon & -2 \end{bmatrix} \; ; \; \mathbf{b} = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \; ; \; \mathbf{c}^T = \begin{bmatrix} 1 & 1 \end{bmatrix}$$

corresponding to the cascading of

$$\frac{1}{s+1} \times \frac{s+2+\epsilon}{s+2}$$

while realization 2 had a structure given by

$$\mathbf{A} = \begin{bmatrix} -2 & 0 \\ 1 & -1 \end{bmatrix} \; ; \; \mathbf{b} = \begin{bmatrix} \epsilon \\ 1 \end{bmatrix} \; ; \; \mathbf{c}^T = \begin{bmatrix} 0 & 1 \end{bmatrix}$$

corresponding to the cascading of

$$\frac{s+2+\epsilon}{s+2} \times \frac{1}{s+1}$$

Only realization 1 was used in this case, and $\bar{\mathbf{A}} = \{a_{ij}\delta_{ij}\}$ was selected. The input was

$$u(t) = \sin(0.2t) + \cos(0.4t)$$

and the results are presented in Table 3.3.

The results are very similar to those seen in standard model order reduction. The states corresponding to large Hankel singular values benefit more from the application of high order discretization, especially for larger sampling periods. As the order of magnitude of the Hankel singular values becomes comparable (as seen in C9-C12), it becomes less critical where the higher order approximations are assigned.

3.9.3 Examination of the Effects of State Space Structure upon Discretization Error

In this case, the importance of state space structure was considered. The choice $\bar{\mathbf{A}} = \{a_{ij}\delta_{ij}\}$ was retained throughout the simulations of this section.

Trials C5-C8 were repeated for the cascade realization 2 referred to in the last section. The difference can be seen particularly clearly by comparing C6 in Table 3.3, to D3 in Table 3.4. This difference can be explained by the fact that large discretization errors generated in the state corresponding to the large Hankel singular value do not propagate through the whole system in D3.

Next a number of state space structures were examined. The system of the last section was used, with $\epsilon = 1 \times 10^{-4}$, and the optimization was done with respect to a simple input

$$u(t) = \cos(0.4t)$$

All integrals were approximated to first order. The results are displayed in Table 3.5. An error is not presented for studies E7-E10 in Table 3.5 due to the fact that a global minimum was difficult to obtain because of the extremely poor numerical properties of the realization. Notice that the input and output balanced realizations are good, as is the cascade realization with the "more important" state second in the series. The optimal state space structure using the theory of Section 3.7 is also included (E11). The selection $\mathbf{\Omega}_u = 1$ and $\mathbf{\Omega}_\eta = 100\mathbf{I}_2$ is used to obtain this realization.

Trial	ϵ	HSV	T	Allocation	Error
C1	0.01	0.5028 0.0003	0.2	1 1	0.0252
C2	0.01	0.5028 0.0003	0.2	1 2	0.0252
C3	0.01	0.5028 0.0003	0.2	2 1	0.0251
C4	0.01	0.5028 0.0003	0.2	2 2	0.0251
C5	0.01	0.5028 0.0003	1	1 1	0.1543
C6	0.01	0.5028 0.0003	1	1 2	0.1502
C7	0.01	0.5028 0.0003	1	2 1	0.1300
C8	0.01	0.5028 0.0003	1	2 2	0.1297
C9	1	0.7854 0.0354	1	1 1	0.2580
C10	1	0.7854 0.0354	1	1 2	0.2403
C11	1	0.7854 0.0354	1	2 1	0.2303
C12	1	0.7854 0.0354	1	2 2	0.2211

Table 3.3: Effect of Hankel singular values

Trial	ϵ	HSV	T	Allocation	Error
D1	0.01	0.0003 0.5028	1	1 1	0.1543
D2	0.01	0.0003 0.5028	1	1 2	0.1292
D3	0.01	0.0003 0.5028	1	2 1	0.1304
D4	0.01	0.0003 0.5028	1	2 2	0.1254

Table 3.4: Importance of state space structure

Trial	Structure	Error
E1	parallel (modal canonical)	0.1602
E2	cascade realization 1	0.1268
E3	cascade realization 2	0.0894
E4	balanced	0.1264
E5	input balanced	0.0956
E6	output balanced	0.0953
E7	controllability	-
E8	observability	-
E9	controller	-
E10	observer	-
E11	optimal	0.0701

Table 3.5: Importance of state space structure

Eigenvalue	Damping	Frequency	HSV
-0.0030+1.0000i	0.0030	1.0000	4.9594×10^{2}
-0.0030+1.0000i	0.0030	1.0000	4.9880×10^{2}
-0.2921	1.0000	0.2921	6.5085×10^{0}
-0.3894	1.0000	0.3894	2.4159×10^{-2}

Table 3.6: System properties

Simulation studies reveal that the state space structure plays the largest role when there is a large divergence between the Hankel singular values. For instance, inequality (3.6.12) suggests that, as $\zeta \to 1$, the discretization errors may become very large for a canonical second order system realized by a real canonical form. However, in practice the state space structure is not overly critical for canonical second order systems with $0 < \zeta < 1$. This can be attributed to the fact that the Hankel singular values for such systems are not widely divergent.

A major point to be made here is that the ease with which the global minimum of the discretization error can be found may depend heavily on the state space structure, even if the value of the global minimum does not. For example, numerical problems are experienced with the algorithm when a second order system is realized with a real canonical form and $\zeta \to 1$.

3.9.4 Case Study

In this section, the techniques outlined in this chapter are applied to the discretization of a fourth order system. The system under consideration is

$$\mathcal{G}(s) = \frac{5.4808s^3 - 3.3826s^2 + 0.3449s + 0.8349}{s^4 + 0.6874s^3 + 1.1178s^2 + 0.6821s + 0.1137}$$

A low order discretization is desired with only one state integral able to be approximated via a second order numerical approximation. Again, the input is a sum of sinusoids with frequencies 0.5 and 0.2 rad/sec. The sampling period is chosen to be $T = 1$ second, and $\bar{\mathbf{A}} = \mathbf{A}$ with a modal canonical form is selected because of numerical considerations.

Table 3.6 summarizes the system properties. Because of the conjugate pair of poles, the selection of $\bar{\mathbf{A}}$ limits the application of second order approximations to either the third or fourth state. This is not necessarily detrimental as seen from the previous

Trial	Discretization method	Cost
F1	Bilinear transformation	8.2285×10^3
F2	Optimized 1 1 1 1	6.1767×10^2
F3	Optimized 2 2 2 2	2.4960×10^1
F4	Optimized 1 1 2 1	6.2350×10^1
F5	Balanced square root truncation	1.2370×10^6
F6	Balanced stochastic truncation	4.1026×10^5

Table 3.7: Results of case study

sections—states which correspond to a damping factor equal to one are often sensitive to the order of the discretization. However it could be argued that, due to the fact that the first two states are poorly damped, they could be very sensitive as well. The natural frequencies of the third and fourth states are similar, but the third state has a large Hankel singular value in comparison to the fourth state. Therefore the third state was chosen for the second order approximation.

The optimization algorithm was run, and the results are summarized in Table 3.7 and Figure 3-10.

The standard bilinear transformation is included for comparison in trial F1. It should be noted that pre-warping does little to improve this result. Trials F2-F4 show the optimized scheme with corresponding orders of approximation. Notice the cost in F4 is still large compared to F3, indicating that the other states would certainly benefit from second order discretization.

In trials F5-F6, all of the state integrals were approximated by second order discretization. In this case, the optimization was performed, followed by a model order reduction in order to bring the discrete system back to fifth order. This clearly shows the efficacy of the method outlined in this chapter. Blindly optimizing with all state integrals approximated to second order and then model order reducing is not an effective method of discretization.

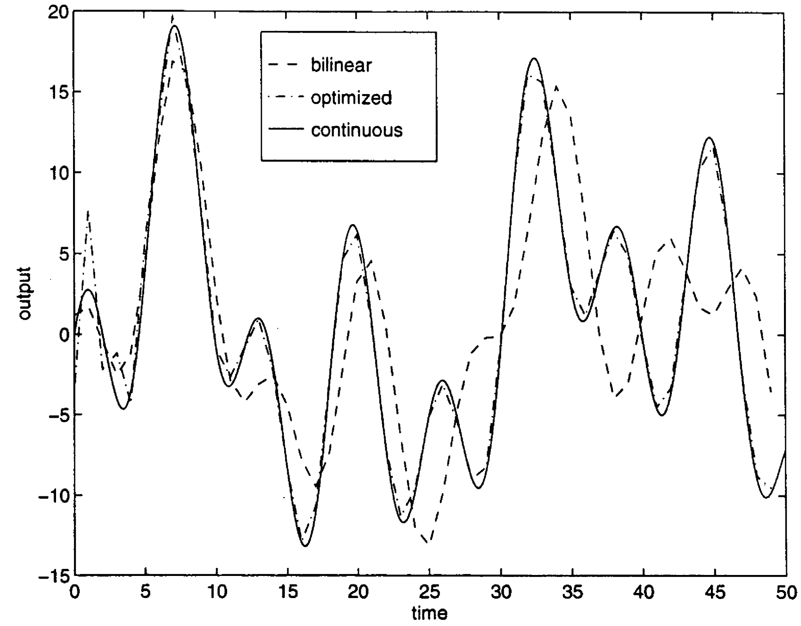

Figure 3-10: Comparison of trials F1 and F4

3.10 Conclusion

This chapter has been concerned with the development of an algorithm for open-loop discretization. The algorithm produces a discretized system of low order, with small discretization error compared to the standard Newton-Cotes schemes. Using an appropriate selection of state space structure and a judicious selection of discretization order for particular states, the discretization error can minimized without resorting to high complexity.

The analysis of this chapter also contributes to the understanding of the factors which effect discretization error and gives some new mathematical and engineering insights. The inherent difficulties of discretization have been captured and the experience and intuition gained is relevant to all forms of discretization, be it open-loop or closed-loop. Consequently, the techniques developed form the basis of a procedure for the digital re-design of analog controllers presented in Chapter 4.

Chapter 4

Closed-Loop Problem

4.1 Introduction

This chapter outlines three methods of digital re-design or closed-loop discretization —the first based on signal invariant transformations and H_2 optimal control theory, the second based on signal invariant transformations and convex optimization theory, and the third based upon an extension of the integral approximation method presented in Chapter 3. In general, each method has a different choice of the closed-loop distance function $\mathfrak{F}_c$ to be used in criterion (2.3.3), introduced in Chapter 2.

The first method involves a two part optimization procedure—a primary optimization and a secondary optimization. The primary optimization solves a H_2 optimal control problem using Youla parameterization, factorization theory and projection mappings. The secondary optimization encompasses some freedom and attempts to find a digital controller with good intersample behaviour. The performance of the resulting digital controller is shown to be very good compared to existing discretization schemes, from both a time-domain and frequency-domain perspective.

The second method, based on convex optimization, is motivated by two factors. The first is the great flexibility that the method allows. A number of different choices of $\mathfrak{F}_c$ can be made and a number of constraints can be incorporated into the design. The second factor is the numerical robustness of convex optimization algorithms. In some cases, the first method suffers from numerical difficulties in finding the optimal Youla parameter $\mathbf{Q}(\cdot)$—this is alleviated using this algorithm.

The third method is motivated by the fact that the first two methods produce a digital controller of higher order than the original analog controller. As a result, some form of controller reduction is generally required. This is not the case with the third method. At the outset of the design, the order of the digital controller is selected. In some applications, this can prove advantageous.

This chapter is organized as follows. Section 4.2 presents the theory of the H_2 optimal control method of controller discretization. The convex optimization method is presented in Section 4.3. The final method based on integral replacement is given in Section 4.4. Simulation results are presented in Section 4.5 with conclusions drawn in Section 4.6.

4.2 H_2 Optimal Control Method

4.2.1 Introduction

In the open-loop case, the use of signal invariant transformations enables the design of a digital system whose output matches perfectly a given analog system at the sample points. Using similar techniques, matching can be achieved in the closed-loop setting. More precisely stated, in response to a given reference signal, it is possible to digitally re-design an analog controller such that the output of the closed-loop sampled-data system matches that of the original analog system, either at the sample points or at any intersample point. However unstable pole-zero cancellations may result from this approach. A technique which avoids this problem is introduced. The methodologies of signal invariant transformations, Youla parameterizations [87, 88] and optimal control theory underpin this method. The method has two parts: a primary optimization and a secondary optimization.

The goal of the primary optimization is to find a digital controller which minimizes the difference between the l_2 norm of the sampled-data system output and that of the analog system at time $kT + \xi$, where T is the sample period, ξ is an intersample point (i.e. $0 \leq \xi < T$), and k is an integer. The optimization is performed in response to a given reference signal.

The need for a secondary optimization arises form the fact that the primary optimization minimizes the l_2 norm at a given ξ, but does not consider other intersample points.

This problem is addressed by re-designing the analog controller based on the optimization at a particular intersample point ξ^*, where ξ^* is selected such that the behaviour at other intersample points is in some way optimized. The choice of a secondary optimization criterion allows some freedom.

As with the approach due to Anderson et al. [1, 43], a digital controller of comparatively high order results with this method, and consequently, standard model reduction methods may need to be employed. A reduction technique for sampled-data systems is given in [56]. As the sampling rate is decreased, the digital controller becomes necessarily of higher order to enable the preservation of performance.

4.2.2 Primary Optimization - l_2 Minimization of Discretization Error at an Intersample Point

In this section, theory is developed which enables matching at a single intersample point. More precisely, a digital controller is found which minimizes the l_2 norm between a closed-loop analog control system and a closed-loop sampled-data system at a single intersample point ξ, $0 \leq \xi < T$; the minimization performed with respect to a given reference signal.

The strictly proper linear time-invariant system $\boldsymbol{\mathcal{R}}(s) \in \mathbb{C}^{n\times 1}$, whose impulse response generates the reference signal $\mathbf{r}(t)$, is assumed to be of the form

$$\boldsymbol{\mathcal{R}}(s) = \begin{bmatrix} r_1(s) & r_2(s) & \cdots & r_n(s) \end{bmatrix}^T \tag{4.2.1}$$

with $\delta(r_i(s)) = 1$, $i = 1, 2, \ldots, n$. This restriction on the relative degree is necessary for the realizability of the digital controller generated from the discretization algorithm—an improper controller may result otherwise. If this condition is not satisfied for a particular $r_i(s)$, then a modification is required. It may be sufficient to alter the phase of a particular $r_i(s)$. For example, if the desired reference trajectory is $r(t) = \sin(\omega t)$ then $\mathcal{R}(s) = \omega(s^2 + \omega^2)^{-1}$ which is unsatisfactory for this method. By changing the reference signal to $r(t) = \sin(\omega t + \frac{\pi}{2}) = \cos(\omega t)$ with $\mathcal{R}(s) = s(s^2+\omega^2)^{-1}$, the difficulties are rectified. Another solution to this problem is to cascade the $r_i(s)$ with an improper low pass filter, i.e. define a new $r_i(s)$ by

$$\tilde{r}_i(s) \triangleq r_i(s) \times \frac{(\omega_c + s)^{\delta(r_i(s))-1}}{\omega_c} \tag{4.2.2}$$

In this case, via a careful selection of ω_c, it is possible to make $\tilde{r}_i(s) \approx r_i(s)$ over the frequency range of interest.

Define $\mathbf{P}^s_T(z)$ and $\mathbf{P}^s_{T,\xi}(z)$ to be the step invariant transformation and ξ-offset step invariant transformations of $\mathcal{P}(s) \in \mathbb{C}^{n\times m}$ respectively (with sampling period T). Let $\mathbf{R}_T(z)$ be the impulse invariant transformation of $\mathcal{R}(s)$. Let the closed-loop transfer function $\mathcal{H}(s)$ of the analog system be defined by

$$\mathcal{H}(s) \triangleq (\mathbf{I}_n + \mathcal{P}(s)\mathcal{C}(s))^{-1}\mathcal{P}(s)\mathcal{C}(s) \tag{4.2.3}$$

and define

$$\mathcal{Y}(s) \triangleq \mathcal{H}(s)\mathcal{R}(s) \tag{4.2.4}$$

Finally, define

$$\mathbf{Y}_{T,\xi}(z) \triangleq Z_{T,\xi}\{\mathcal{Y}(s)\} \tag{4.2.5}$$

which can be written as

$$\mathbf{Y}_{T,\xi}(z) = \mathbf{H}^r_{T,\xi}(z)\mathbf{R}_T(z) \tag{4.2.6}$$

where $\mathbf{H}^r_{T,\xi}(z)$ is the ξ-offset signal invariant transformation of $\mathcal{H}(s)$ with respect to $\mathbf{r}(t)$.

With these definitions, the impulse response of $\mathbf{Y}_{T,\xi}(z)$ generates the sequence $\mathbf{y}(kT+\xi)$, $k = 0, 1, \ldots, \infty$. Moreover, the impulse response of the system formed by

$$\mathbf{P}^s_{T,\xi}(z)(\mathbf{I}_n + \mathbf{C}_d(z)\mathbf{P}^s_T(z))^{-1}\mathbf{C}_d(z)\mathbf{R}_T(z) \tag{4.2.7}$$

generates the sequence $\hat{\mathbf{y}}(kT+\xi)$, $k = 0, 1, \ldots, \infty$.

Introduce a cost function $J(\mathbf{C}_d(z), \xi)$ where

$$\begin{aligned} J(\mathbf{C}_d(z), \xi) &= \|\hat{\mathbf{y}}(kT+\xi) - \mathbf{y}(kT+\xi)\|_2^2 \text{ for some } 0 \leq \xi < T && (4.2.8) \\ &= \|\{\mathbf{P}^s_{T,\xi}(z)(\mathbf{I}_n + \mathbf{C}_d(z)\mathbf{P}^s_T(z))^{-1}\mathbf{C}_d(z) - \mathbf{H}^r_{T,\xi}(z)\}\mathbf{R}_T(z)\|_2^2 && (4.2.9) \end{aligned}$$

A stabilizing digital controller $\mathbf{C}_d(z)$ is sought which minimizes (4.2.9). Before giving the solution, a number of observations will be made. In Figure 4-1, a diagrammatic representation of the cost function $J(\mathbf{C}_d(z), \xi)$ is shown. The diagram shows that the states of $\mathbf{H}^r_{T,\xi}(z)$ and $\mathbf{R}_T(z)$ are uncontrollable from $\mathbf{C}_d(z)$. The assumption of stability in the analog closed-loop system ensures the stabilizability of the states of $\mathbf{H}^r_{T,\xi}(z)$. Furthermore, if the reference model is stable then the entire system is stabilizable. However, if this is not the case, for example when the reference model has poles on the unit circle, then the problem is not as straightforward.

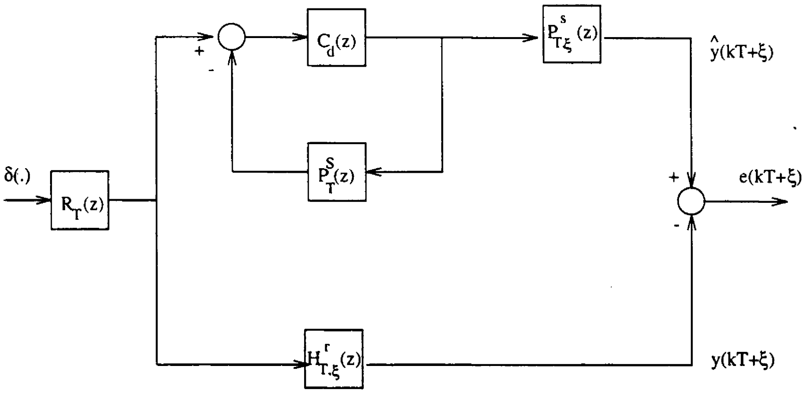

Figure 4-1: Block diagram representation of cost function

Consider the case when $\xi = 0$. In this case, the block diagram of Figure 4-1 simplifies in that the top branch reduces to the standard unity feedback arrangement with $\mathbf{C}_d(z)$ in series with $\mathbf{P}_T^s(z)$. Let the state space representation of $\mathbf{R}_T(z)$ be

$$\mathbf{R}_T(z) \equiv [\mathbf{Q}_d, \mathbf{R}_d, \mathbf{S}_d, \mathbf{T}_d] \tag{4.2.10}$$

The signal $\mathbf{r}(kT)$ is generated by the impulse response of $\mathbf{R}_T(z)$. Suppose the step invariant transformation of the plant, $\mathbf{P}_T^s(z)$ has a state space representation,

$$\mathbf{P}_T^s(z) \equiv [\mathbf{F}, \mathbf{G}, \mathbf{H}, \mathbf{J}] \tag{4.2.11}$$

It is known [30], that a necessary condition for perfect reference tracking is that each mode present in $\mathbf{r}(kT)$ must also be present in the open-loop discrete model of the plant. That is, the eigenvalues of $\mathbf{Q}_d$ in (4.2.10) must form a subset of the eigenvalues of $\mathbf{F}$. If this is not the case, then those eigenvalues that are not present must be supplied by $\mathbf{C}_d(z)$.

Specifically, assume without loss of generality that the matrix $\mathbf{Q}_d$ in (4.2.10) is in block diagonal form

$$\mathbf{Q}_d = \text{block diagonal}\{\mathbf{Q}_1, \mathbf{Q}_2\},$$

where all eigenvalues $\lambda_j(\mathbf{Q}_1)$ of $\mathbf{Q}_1$ and $\lambda_j(\mathbf{Q}_2)$ of $\mathbf{Q}_2$ satisfy the respective conditions

$$|\lambda_j(\mathbf{Q}_1)| < 1 \quad ; \quad |\lambda_i(\mathbf{Q}_2)| \geq 1.$$

For asymptotic tracking, only the eigenvalues of $\mathbf{Q}_2$ need to be included in $\mathbf{P}_T^s(z)\mathbf{C}_d(z)$. Given that this condition is satisfied and $\mathbf{C}_d(z)$ is a stabilizing controller for the top branch, then $\hat{\mathbf{y}}(kT) \to \mathbf{r}(kT)$ as $k \to \infty$. Since $\mathbf{y}(kT) \to \mathbf{r}(kT)$ asymptotically by nature of the design of $\mathcal{H}(s)$, then $\mathbf{e}(kT) \to \mathbf{0}$ as $k \to \infty$. Given that the rate of convergence is exponential, this suggests that it is possible to achieve a finite cost function (4.2.9), even for unstable reference models. A variation of this argument can be made for $\xi \neq 0$.

Stable $\mathcal{R}(s)$

Let a doubly coprime factorization of the step invariant transformation of $\mathcal{P}(s)$ be

$$\mathbf{P}^{s}_{T}(z) = \mathbf{N}_R(z)\mathbf{D}_R^{-1}(z) = \mathbf{D}_L^{-1}(z)\mathbf{N}_L(z) \tag{4.2.12}$$

with

$$\begin{bmatrix} \mathbf{V}_R(z) & \mathbf{U}_R(z) \\ -\mathbf{N}_L(z) & \mathbf{D}_L(z) \end{bmatrix} \begin{bmatrix} \mathbf{D}_R(z) & -\mathbf{U}_L(z) \\ \mathbf{N}_R(z) & \mathbf{V}_L(z) \end{bmatrix} = \begin{bmatrix} \mathbf{I}_m & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_n \end{bmatrix} \tag{4.2.13}$$

Let

$$\mathbf{P}^{s}_{T,\xi}(z) = \mathbf{N}_\xi(z)\mathbf{D}_R^{-1}(z) \tag{4.2.14}$$

Define quantities $\mathbf{T}_1(z), \mathbf{T}_2(z), \mathbf{T}_3(z)$ by

$$\mathbf{T}_1(z) \triangleq \mathbf{N}_\xi(z)\mathbf{U}_R(z)\mathbf{R}_T(z) - \mathbf{H}^{r}_{T,\xi}(z)\mathbf{R}_T(z) \in \mathbb{C}^{n\times 1} \tag{4.2.15}$$

$$\mathbf{T}_2(z) \triangleq \mathbf{N}_\xi(z) \in \mathbb{C}^{n\times m} \tag{4.2.16}$$

$$\mathbf{T}_3(z) \triangleq \mathbf{D}_L(z)\mathbf{R}_T(z) \in \mathbb{C}^{n\times 1} \tag{4.2.17}$$

Theorem 4.1 *(See Figure 2-1) Consider an asymptotically stable analog unity feedback control system where the strictly proper plant $\mathcal{P}(s) \in \mathbb{C}^{n\times m}$, the strictly proper stable reference model $\mathcal{R}(s) \in \mathbb{C}^{n\times 1}$, and the proper analog controller $\mathcal{C}(s) \in \mathbb{C}^{m\times n}$ are all given.*

Consider also a sampled-data unity feedback control system with plant $\mathcal{P}(s)$, reference model $\mathcal{R}(s)$ and digital controller $\mathbf{C}_d(z)$ with zero order hold. The internally stabilizing digital controller $\mathbf{C}_d^*(z)$ which minimizes the cost function (4.2.9) is given by

$$\mathbf{C}_d^*(z) = (\mathbf{V}_R(z) - \mathbf{Q}^*(z)\mathbf{N}_L(z))^{-1}(\mathbf{U}_R(z) + \mathbf{Q}^*(z)\mathbf{D}_L(z)) \tag{4.2.18}$$

$\mathbf{Q}^*(z)$ is given by

$$\mathbf{Q}^*(z) = -\mathbf{T}_{2_o}^{-1}(z)\Pi^- \left[\mathbf{T}_{2_i}^{\sim}(z)\mathbf{T}_1(z)\mathbf{T}_{3_{ci}}^{\sim}(z)\right]\mathbf{T}_{3_{co}}^{-1}(z) \in \mathbb{C}^{m\times n} \tag{4.2.19}$$

where the quantity $\mathbf{T}_1(z)$ is defined in equation (4.2.15). The quantities $\mathbf{T}_2(z)$ and $\mathbf{T}_3(z)$ defined in equations (4.2.16) and (4.2.17) have been factored by an inner-outer factorization on $\mathbf{T}_2(z)$ and a co-inner-co-outer factorization on $\mathbf{T}_3(z)$, i.e.

$$\mathbf{T}_2(z) = \mathbf{T}_{2_i}(z)\mathbf{T}_{2_o}(z) \tag{4.2.20}$$

$$\mathbf{T}_3(z) = \mathbf{T}_{3_{co}}(z)\mathbf{T}_{3_{ci}}(z) \tag{4.2.21}$$

The optimal value of (4.2.9) is given by

$$\begin{aligned}&\|\Pi^{+}[\mathbf{T}_{2_i}^{\sim}(z)\mathbf{T}_1(z)\mathbf{T}_{3_{ci}}^{\sim}(z)]\|_2^2+\|\mathbf{T}_{2_i}^{\sim}(z)\mathbf{T}_1(z)(\mathbf{I}_1-\mathbf{T}_{3_{ci}}^{\sim}(z)\mathbf{T}_{3_{ci}}(z))\|_2^2\\&+\|(\mathbf{I}_n-\mathbf{T}_{2_i}(z)\mathbf{T}_{2_i}^{\sim}(z))\mathbf{T}_1(z)\mathbf{T}_{3_{ci}}^{\sim}(z)\|_2^2+\|(\mathbf{I}_n-\mathbf{T}_{2_i}(z)\mathbf{T}_{2_i}^{\sim}(z))\mathbf{T}_1(z)(\mathbf{I}_1-\mathbf{T}_{3_{ci}}^{\sim}(z)\mathbf{T}_{3_{ci}})\|_2^2\end{aligned} \tag{4.2.22}$$

Proof: Assuming the standard stabilizability and detectability assumptions on the continuous plant $\mathcal{P}(s)$, then provided a non-pathological sampling period (i.e. the discretization of the loop does not introduce unstabilizable or undetectable modes), $P_T^s(z)$ is both stabilizable and detectable (see [29]). From the Youla parameterization theory of [87, 88], given coprime factorizations of $\mathbf{P}_T^s(z)$ as described by the equations (4.2.12) and (4.2.13), the class of all stabilizing controllers for $\mathbf{P}_T^s(z)$ is described by

$$\mathbf{C}_d(z)=(\mathbf{V}_R(z)-\mathbf{Q}(z)\mathbf{N}_L(z))^{-1}(\mathbf{U}_R(z)+\mathbf{Q}(z)\mathbf{D}_L(z)),\ \ \mathbf{Q}(z)\in RH^{\infty} \tag{4.2.23}$$

From the stability theory of sampled-data systems presented in [29], we know that this $\mathbf{C}_d(z)$ also stabilizes the sampled-data system (under the same non-pathological sampling period condition). Substituting (4.2.23), (4.2.12) and (4.2.14) into the cost function expression (4.2.9) and simplifying gives

$$\begin{aligned}J(\mathbf{C}_d(z),\xi)&\equiv J(\mathbf{Q}(z),\xi)\\&=\|\mathbf{N}_{\xi}(z)\mathbf{U}_R(z)\mathbf{R}_T(z)-\mathbf{H}_{T,\xi}^{r}(z)\mathbf{R}_T(z)+\mathbf{N}_{\xi}(z)\mathbf{Q}(z)\mathbf{D}_L(z)\mathbf{R}_T(z)\|_2^2\\&=\|\mathbf{T}_1(z)+\mathbf{T}_2(z)\mathbf{Q}(z)\mathbf{T}_3(z)\|_2^2\end{aligned} \tag{4.2.24}$$

where the quantities $\mathbf{T}_1(z),\mathbf{T}_2(z),\mathbf{T}_3(z)$ are defined according to equations (4.2.15), (4.2.16), and (4.2.17). Perform an inner-outer factorization on $\mathbf{T}_2(z)$ and a co-inner-co-outer factorization (see Theorem A.1) on $\mathbf{T}_3(z)$ according to equations (4.2.20) and (4.2.20) and apply a unitary transformation, i.e.

$$\begin{aligned}J(\mathbf{Q}(z),\xi)&=\|\mathbf{T}_1(z)+\mathbf{T}_{2_i}(z)\mathbf{T}_{2_o}(z)\mathbf{Q}(z)\mathbf{T}_{3_{co}}(z)\mathbf{T}_{3_{ci}}(z)\|_2^2\\&=\left\|\begin{bmatrix}\mathbf{T}_{2_i}^{\sim}(z)\\ \mathbf{I}_n-\mathbf{T}_{2_i}(z)\mathbf{T}_{2_i}^{\sim}(z)\end{bmatrix}\{\mathbf{T}_1(z)+\mathbf{T}_{2_i}(z)\mathbf{T}_{2_o}(z)\mathbf{Q}(z)\mathbf{T}_{3_{co}}(z)\mathbf{T}_{3_{ci}}(z)\}\right.\\&\qquad\left.\times\begin{bmatrix}\mathbf{T}_{3_{ci}}^{\sim}(z) & \mathbf{I}_1-\mathbf{T}_{3_{ci}}^{\sim}(z)\mathbf{T}_{3_{ci}}(z)\end{bmatrix}\right\|_2^2\\&=\|\mathbf{T}_{2_i}^{\sim}(z)\mathbf{T}_1(z)\mathbf{T}_{3_{ci}}^{\sim}(z)+\mathbf{T}_{2_o}(z)\mathbf{Q}(z)\mathbf{T}_{3_{co}}(z)\|_2^2\\&\quad+\|\mathbf{T}_{2_i}^{\sim}(z)\mathbf{T}_1(z)(\mathbf{I}_1-\mathbf{T}_{3_{ci}}^{\sim}(z)\mathbf{T}_{3_{ci}}(z))\|_2^2\\&\quad+\|(\mathbf{I}_n-\mathbf{T}_{2_i}(z)\mathbf{T}_{2_i}^{\sim}(z))\mathbf{T}_1(z)\mathbf{T}_{3_{ci}}^{\sim}(z)\|_2^2\\&\quad+\|(\mathbf{I}_n-\mathbf{T}_{2_i}(z)\mathbf{T}_{2_i}^{\sim}(z))\mathbf{T}_1(z)(\mathbf{I}_1-\mathbf{T}_{3_{ci}}^{\sim}(z)\mathbf{T}_{3_{ci}})\|_2^2\end{aligned} \tag{4.2.25}$$

Only the first term depends on $\mathbf{Q}(z)$ and can be written as

$$\|\Pi^{-}[\mathbf{T}_{2_i}^{\sim}(z)\mathbf{T}_1(z)\mathbf{T}_{3_{ci}}^{\sim}(z)] + \mathbf{T}_{2_o}(z)\mathbf{Q}(z)\mathbf{T}_{3_{co}}(z)\|_2^2 + \|\Pi^{+}[\mathbf{T}_{2_i}^{\sim}(z)\mathbf{T}_1(z)\mathbf{T}_{3_{ci}}^{\sim}(z)]\|_2^2$$

The minimizing $\mathbf{Q}(z) \in RH^{\infty}$ is then given by equation (4.2.19) and the optimal cost (4.2.22) follows. ∎

The case of perfect matching is of particular interest, i.e. the case when the cost function (4.2.9) can be made identically equal to zero. From equation (4.2.24), this is only possible when $\mathbf{T}_2(z) \in \mathbb{C}^{n \times m}$ and $\mathbf{T}_3(z) \in \mathbb{C}^{n \times 1}$ have stable inverses. Clearly this can only occur for SISO systems. This gives the following result:

Corollary 4.1 *Necessary conditions for the cost function (4.2.9) to be made identically zero are:*

1. *The plant $\mathcal{P}(s)$ is asymptotically stable*
2. *$P_{T,\xi}^{s}(z)$ is minimum phase for $\xi \in [0, T)$*
3. *$R_T(z)$ has a stable inverse.*

The following result is for SISO systems in the case $\xi = 0$.

Corollary 4.2 *Given $\delta(\mathcal{R}(s)) = 1$ and $P_T^s(z)$ both asymptotically stable and of minimum phase then*

$$C_d^*(z) = \frac{Z_T\{\mathcal{P}(s)\mathcal{C}(s)\mathcal{E}(s)\}}{Z_T\{\mathcal{E}(s)\}P_T^s(z)} \tag{4.2.26}$$

where $P_T^s(z)$ is the step invariant transformation of the plant $\mathcal{P}(s)$ and

$$\mathcal{E}(s) = \frac{\mathcal{R}(s)}{1 + \mathcal{P}(s)\mathcal{C}(s)} \tag{4.2.27}$$

For zero initial conditions and any impulse applied at $t = 0$, we have

$$\hat{y}(kT) = y(kT)\ ; \quad \textit{integer } k \geq 0$$

In particular, for almost all sampling periods:

(i) $C_d(z)$ is of order $n_r + n_p + n_c$ where $\mathcal{P}(s)$ and $\mathcal{C}(s)$ are of order n_p and n_c poles respectively, and $\mathcal{R}(s)$ has n_r poles not contained in $\mathcal{P}(s)\mathcal{C}(s)$.

(ii) $C_d(z)$ is proper with $\delta(C_d(z)) = 0$.

(iii) If $\mathcal{P}(s)$ is asymptotically stable and $\mathcal{E}(s)$ is of minimum phase, then $P_T^s(z)C(z)$ is asymptotically stable for T sufficiently large.

(iv) If both $\mathcal{P}(s)\mathcal{C}(s)$ and $\mathcal{E}(s)$ are of minimum phase, then $P_T^s(z)C_d(z)$ is of minimum phase for T sufficiently large.

Proof: The results follow by straightforward algebraic manipulations. Property (ii) follows from previous arguments and properties (iii)-(iv) follow directly from Theorem 3.1. ∎

Unstable $\mathcal{R}(s)$

In this section, a method is outlined which deals with an unstable reference model (i.e. with poles on the $j\omega$ axis or in the right-half plane). A variation of the preceeding theory is used. Redefine the quantities $\mathbf{T}_1(z)$, $\mathbf{T}_2(z)$ and $\mathbf{T}_3(z)$ (first defined in (4.2.15), (4.2.16), and (4.2.17) respectively) as

$$\mathbf{T}_1(z) \triangleq \mathbf{N}_\xi(z)\mathbf{U}_R(z) - \mathbf{H}_{T,\xi}^r(z) \in \mathbb{C}^{n\times n} \tag{4.2.28}$$

$$\mathbf{T}_2(z) \triangleq \mathbf{N}_\xi(z) \in \mathbb{C}^{n\times m} \tag{4.2.29}$$

$$\mathbf{T}_3(z) \triangleq \mathbf{D}_L(z) \in \mathbb{C}^{n\times n} \tag{4.2.30}$$

Then the minimization of $J(\mathbf{C}_d(z), \xi)$ is equivalent to the minimization of

$$J(\mathbf{Q}(z), \xi) = \|\{\mathbf{T}_1(z) + \mathbf{T}_2(z)\mathbf{Q}(z)\mathbf{T}_3(z)\}\mathbf{R}_T(z)\|_2^2 \tag{4.2.31}$$

with $\mathbf{Q}(z) \in \mathbb{C}^{m\times n}$ and $\mathbf{Q}(z) \in RH^\infty$ (c.f. (4.2.24)). Given the structure of $\mathcal{R}(s)$ (c.f. equation (4.2.1)), $\mathbf{R}_T(z)$ has the form

$$\mathbf{R}_T(z) = \begin{bmatrix} \bar{r}_1(z) & \bar{r}_2(z) & \dots & \bar{r}_n(z) \end{bmatrix}^T \text{ where } \bar{r}_i(z) = Z_T\{r_i(s)\},\ i = 1, 2, \dots, n \tag{4.2.32}$$

A necessary condition for a finite cost in (4.2.31) is that any unstable poles in the $\bar{r}_i(z)$ are cancelled by corresponding transmission zeros in $\mathbf{T}_1(z) + \mathbf{T}_2(z)\mathbf{Q}(z)\mathbf{T}_3(z)$. These zeros must be positioned by imposing interpolation constraints on $\mathbf{Q}(z)$.

If $r_i(z)$ has an unstable pole (for simplicity assumed to be a simple pole) for some i, $0 \leq i \leq n$ at $z = z_j$, then a sufficient condition for the pole-zero cancellation is that $\mathbf{T}_1(z) + \mathbf{T}_2(z)\mathbf{Q}(z)\mathbf{T}_3(z)$ is identically equal to zero for all elements in the i^{th} column when $z = z_j$, i.e.

$$\mathbf{T}_1^{[i]}(z_j) + \mathbf{T}_2(z_j)\mathbf{Q}(z_j)\mathbf{T}_3^{[i]}(z_j) = \mathbf{0} \qquad (4.2.33)$$

where $\mathbf{T}_k^{[i]}(\cdot)$ means the i^{th} column of $\mathbf{T}_k(\cdot)$. The problem is now to find all stable $\mathbf{Q}(z)$ that satisfy equation (4.2.33).

The problem given by equation (4.2.33) is related to the tangential interpolation problems discussed in [7, 8]. The tangential interpolation problem involves finding all stable $\mathbf{W}(z) \in \mathbb{C}^{n \times m}$ such that

$$\mathbf{x}^T(z_k)\mathbf{W}(z_k) = \mathbf{y}^T(z_k), \; k = 1, \ldots, n \qquad (4.2.34)$$

where $\mathbf{x}(z) \in \mathbb{C}^{n \times 1}$, $\mathbf{y}(z) \in \mathbb{C}^{m \times 1}$. An elegant solution exists to this problem. However, the problem (4.2.33) with matrix functions on both sides of the $\mathbf{Q}(\cdot)$ has not been well studied.

The following lemma provides a parameterization of all solutions to (4.2.33).

Lemma 4.1 *Let $\mathbf{Q}(z) = \hat{\mathbf{Q}}(z) \in RH^\infty$ be any solution to equation (4.2.33). Then provided*

$$\mathbf{T}_2(z_j)[\mathbf{T}_2(z_j)]^\# \mathbf{T}_1^{[i]}(z_j)[\mathbf{T}_3^{[i]}(z_j)]^\# \mathbf{T}_3^{[i]}(z_j) = \mathbf{T}_1^{[i]}(z_j) \qquad (4.2.35)$$

then $\hat{\mathbf{Q}}(z)$ can be written as

$$\begin{aligned}\hat{\mathbf{Q}}(z) &= -[\mathbf{T}_2(z_j)]^\# \mathbf{T}_1^{[i]}(z_j)[\mathbf{T}_3^{[i]}(z_j)]^\# - \mathbf{T}_2(z_j)[\mathbf{T}_2(z_j)]^\# \mathbf{M}_j^F [\mathbf{T}_3^{[i]}(z_j)]^\# \mathbf{T}_3^{[i]}(z_j) \\ &\quad + \mathbf{M}_j^F + \mathbf{Q}_j^F(z)\frac{z - z_j}{z - \zeta_j} \qquad (4.2.36)\end{aligned}$$

where $0 < \zeta_j < 1$, $\mathbf{M}_j^F$ is any $m \times n$ matrix, and $\mathbf{Q}_j^F(z)$ is any $m \times n$ rational transfer matrix in RH^∞.

Proof: Follows from standard linear algebra, see for example [84]. ∎

The disadvantage with this particular parameterization is that it is composed of a free rational transfer matrix and a free constant matrix. This is in contrast to the problem given by equation (4.2.34), where the solution can be written as an affine function of a free rational transfer matrix only. Nevertheless, this lemma enables a series of interpolation constraints to be imposed according to the unstable poles of the reference model. The problem then involves an optimization over the free parameters $\mathbf{Q}_j^F(z)$ and $\mathbf{M}_j^F$. The optimization is in fact convex.

To demonstrate the method assume that only one constraint is present at $z = z_1$. The cost criterion (4.2.31) can be written as

$$J(\mathbf{Q}(z), \xi) = \|\tilde{\mathbf{T}}_1(z) + \tilde{\mathbf{T}}_2(z)\mathbf{Q}_1^F(z)\tilde{\mathbf{T}}_3(z)\|_2^2 \tag{4.2.37}$$

with

$$\begin{aligned}
\tilde{\mathbf{T}}_1(z) &\triangleq \mathbf{T}_1(z) + \mathbf{T}_2(z)\{-[\mathbf{T}_2(z_1)]^{\#}\mathbf{T}_1^{[i]}(z_1)[\mathbf{T}_3^{[i]}(z_1)]^{\#} \\
&\quad -\mathbf{T}_2(z_1)[\mathbf{T}_2(z_1)]^{\#}\mathbf{M}_1^F[\mathbf{T}_3^{[i]}(z_1)]^{\#}\mathbf{T}_3^{[i]}(z_1) + \mathbf{M}_j^F\}\mathbf{R}_T(z) && (4.2.38) \\
\tilde{\mathbf{T}}_2(z) &\triangleq \mathbf{T}_2(z) && (4.2.39) \\
\tilde{\mathbf{T}}_3(z) &\triangleq \frac{z - z_1}{z - \zeta_1}\mathbf{T}_3(z)\mathbf{R}_T(z) && (4.2.40)
\end{aligned}$$

The optimization is then performed as in the stable case by first finding an expression for the the optimal $\mathbf{Q}_1^F(z)$ in terms of $\mathbf{M}_1^F$. The resulting cost (c.f. equation (4.2.22)) is then given in terms of $\mathbf{M}_1^F$ and so a secondary minimization can be performed with respect to this parameter. The optimal $\mathbf{M}_1^F$ found can be substituted back into the expression for $\mathbf{Q}_1^F(z)$.

It should be pointed out that the above theory is greatly simplified in the SISO case, with equation (4.2.36) simplifying to

$$\hat{Q}(z) = -\frac{T_1(z_j)}{T_2(z_j)T_3(z_j)} + Q_j^F(z)\frac{z - z_j}{z - \zeta_j} \tag{4.2.41}$$

Further discussion of the unstable reference model case appears in the next section.

4.2.3 Secondary Optimization and Main Algorithm

To this point, techniques have been developed to enable the l_2 minimization of the difference between the output of analog system and the sampled-data system, with respect to a given reference signal, at a given intersample point. The question now posed is whether a digital controller can be found such that the behaviour of the analog and sampled-data systems can be made "close" in some sense, over the entire intersample period.

It must be kept in mind that the sampled-data system operates in open-loop between the sample points (given the zero-order hold), so there are fundamental limits imposed upon the success of matching for a given reference signal class. It is well known [6] that for a general continuous-time reference signal, a sampled-data system may exhibit non-zero steady state tracking error, despite having zero tracking error at the sample points. This phenomenom, called hidden oscillations or simply ripple, is now discussed.

Ripple is defined as error between sampling instants when there is zero error at the sampling instants. Ripple may have a constant amplitude or it may decrease with time. It is even possible to demonstrate that a system can have zero error at sampling instants yet have ripple with growing amplitude.

It is known that if the steady state of either the reference input or the disturbance is not constant (e.g. ramps, sinusoids, polynomials, e.t.c.), then a sampled-data controller using zero-order hold will give rise to ripple error. That is, intersampling error exists and does not decay, although the steady-state error is zero at the sampling instances. As discussed in [6], ripple can also occur if the observability of the open-loop system is lost due to sampling or if the unobservable open-loop modes due to sampling are oscillatory. Ripple may also occur if poorly damped process zeros are cancelled by the regulator.

To eliminate intersample ripple, two methods are generally employed:

1. A higher order hold function may be employed. An example of this approach can be found in [77].

2. A continuous-time internal model of the reference model can be inserted into the control loop. In [31] an internal model principle for sampled-data systems is given. In a sampled-data system, the necessary and sufficient conditions for the continuous response to be ripple free are that the continuous plant is controllable with discrete inputs at period T and that the combination of plant plus hold plus compensator must have a continuous internal model of the exogenous input, that is observable from the output.

The significance of the second method is that it presents an alternative to the theory presented in Section 4.2.2. The sampled-data internal model principle means that the interpolation conditions (c.f. equation (4.2.33)) are automatically satisfied. However the designer must decide upon the relative merits of a design with some ripple, as may be achieved using the theory of Section 4.2.2, or a design which is ripple free, obtained by using a sampled-data internal model.

With these thoughts in mind, the following discretization scheme is proposed:

1. Given an intersample point ξ, a digital controller is found using the method developed in Section 4.2.2, such that the "primary" cost function (the l_2 norm) is minimized at ξ. Define $\mathbf{C}^*_{d,\xi}(z)$ as

$$\begin{aligned}\mathbf{C}^*_{d,\xi}(z) &\triangleq \arg\min_{\mathbf{C}_d(z)} ||\hat{\mathbf{y}}(kT+\xi) - \mathbf{y}(kT+\xi)||_2^2 \quad 0 \leq \xi < T && (4.2.42)\\ &= \arg\min_{\mathbf{C}_d(z)} ||\mathbf{P}^s_{T,\xi}(\mathbf{I}_n + \mathbf{P}^s_T(z)\mathbf{C}_d(z))^{-1}\mathbf{C}_d(z)\mathbf{R}_T(z) - \mathbf{Y}_{T,\xi}(z)||_2^2 && (4.2.43)\end{aligned}$$

2. Given $\mathbf{C}^*_{d,\xi}(z)$, a "secondary" cost function is then calculated which reflects the discretization performance over the whole intersample period $[0, T)$. The choice of this "secondary" norm will be discussed.

3. The controller which corresponds to the ξ which minimizes this secondary norm is chosen as the "optimal" digital controller.

4. It may be desirable for the resulting digital controller to be reduced in order via a number of methods such as balanced truncations or the method of [56].

The selection of the secondary cost function encompasses some freedom. This freedom gives the designer additional control over the final sampled-data controller. Four possible choices of secondary cost function are:

1.
$$\Phi_1(\mathbf{C}^*_{d,\xi}(z)) = \sum_{i=0}^{N-1} ||\hat{\mathbf{y}}(kT + \frac{iT}{N}) - \mathbf{y}(kT + \frac{iT}{N})||_2^2 \qquad (4.2.44)$$

This choice of secondary cost function is appropriate only if conditions are met so that the intersample ripple tends to zero as time goes to infinity (see previous discussion). Unless these conditions are met, given a primary optimization at a point ξ which may have a zero steady state tracking error at ξ, the error may be non zero at other intersample points. If this is the case, the secondary cost function will not be finite (a possible solution to this is to evaluate a truncated l_2 norm by summing only a finite number of terms in the time-domain response).

2.

$$\Phi_2(\mathbf{C}^*_{d,\xi}(z)) = \sup_{\beta} \|\hat{\mathbf{y}}(kT+\beta) - \mathbf{y}(kT+\beta)\|_2^2 \tag{4.2.45}$$

This selection requires the same conditions as the first.

3.

$$\Phi_3(\mathbf{C}^*_{d,\xi}(z)) = \sup_{\beta} \sup_{k} \|\hat{\mathbf{y}}(kT+\beta) - \mathbf{y}(kT+\beta)\|_2^2 \tag{4.2.46}$$

This secondary cost criterion can be used in the case of non zero ripple.

4.

$$\Phi_4(\mathbf{C}^*_{d,\xi}(z)) = \sup_{\beta} \limsup_{k} \|\hat{\mathbf{y}}(kT+\beta) - \mathbf{y}(kT+\beta)\|_2^2 \tag{4.2.47}$$

This criterion can be used for looking at steady state performance.

Obviously there could be other appropriate choices of secondary cost criterion.

The first two choices of cost criterion are generally preferable because of the ease of computation (the last two are difficult to calculate). Because of this reason, algorithms which use these are now developed. These are the algorithms which are used in the simulation results which follow. Reiterating, these may by used in the case when conditions that guarantee zero steady state ripple are met.

Both algorithms require the ability to calculate the quantity

$$\Phi(\mathbf{C}^*_{d,\xi}(z), \beta) = \|\hat{\mathbf{y}}(kT+\beta) - \mathbf{y}(kT+\beta)\|_2^2 \tag{4.2.48}$$

for all points $\beta \in [0, T)$. Furthermore the calculation of the partial derivatives of Φ with respect to β are required for the calculation of $\Phi_2(\cdot)$. As with the evaluation of the cost function in Chapter 3, the method found to be most effective for the evaluation of (4.2.48) is the doubling algorithm (c.f. Algorithm 3.1). The method proceeds as follows.

For a given $\mathbf{C}^*_{d,\xi}(z)$ let a state space realization of $(\mathbf{I}_n + \mathbf{C}^*_{d,\xi}(z)\mathbf{P}^s_T(z))^{-1}\mathbf{C}^*_{d,\xi}(z)$ be

$$[\hat{\mathbf{F}}, \hat{\mathbf{G}}, \hat{\mathbf{H}}, \hat{\mathbf{J}}]$$

Notice that this is independent of β. Further, let a state space realization of $\mathbf{P}^s_{T,\beta}(z)$ be

$$[\mathbf{F}_P, \mathbf{G}_P, \mathbf{H}_P(\beta), \mathbf{J}_P(\beta)]$$

Notice that $\mathbf{H}_P$ and $\mathbf{J}_P$ depend upon β. Recall that this realization can be given by (3.2.21). Finally let a state space realization of $\mathbf{Y}_{T,\beta}(z)$ be

$$[\mathbf{F}_Y, \mathbf{G}_Y, \mathbf{H}_Y(\beta), \mathbf{J}_Y(\beta)]$$

with these quantities obtainable via equation (3.2.8). A realization of the system

$$\mathbf{P}^s_{T,\beta}(z)(\mathbf{I}_n + \mathbf{C}^*_{d,\xi}(z)\mathbf{P}^s_T(z))^{-1}\mathbf{C}^*_{d,\xi}(z) - \mathbf{Y}_{T,\beta}(z)$$

is then given by

$$\left[\begin{bmatrix} \mathbf{F}_P & \mathbf{G}_P\hat{\mathbf{H}} & \mathbf{0} \\ \mathbf{0} & \hat{\mathbf{F}} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{F}_Y \end{bmatrix}, \begin{bmatrix} \mathbf{G}_P\hat{\mathbf{J}} \\ \hat{\mathbf{G}} \\ \mathbf{G}_Y \end{bmatrix}, \begin{bmatrix} \mathbf{H}_P(\beta) & \mathbf{J}_P(\beta)\hat{\mathbf{H}} & -\mathbf{H}_Y(\beta) \end{bmatrix}, \mathbf{J}_P(\beta)\hat{\mathbf{J}} - \mathbf{J}_Y(\beta)\right]$$
$$\triangleq \left[\tilde{\mathbf{F}}, \tilde{\mathbf{G}}, \tilde{\mathbf{H}}(\beta), \tilde{\mathbf{J}}(\beta)\right] \tag{4.2.49}$$

The quantity $\Phi(\mathbf{C}^*_{d,\xi}(z), \beta)$ is evaluated numerically via

$$X_k = \sum_{i=0}^{N} \tilde{\mathbf{F}}^i \tilde{\mathbf{G}}\tilde{\mathbf{G}}^T (\tilde{\mathbf{F}}^T)^i, \quad N = 2^k - 1$$

which is evaluated using the doubling algorithm. This in turn gives

$$\Phi(\mathbf{C}^*_{d,\xi}(z), \beta) = \mathrm{tr}\left(\tilde{\mathbf{H}}(\beta)\mathbf{X}_k\tilde{\mathbf{H}}^T(\beta) + \tilde{\mathbf{J}}(\beta)\tilde{\mathbf{J}}^T(\beta)\right) \text{ as } k \to \infty \tag{4.2.50}$$

Typically $k \approx 15$ is adequate. The partial derivatives of $\Phi(\mathbf{C}^*_{d,\xi}(z), \beta)$ with respect to β can also be calculated quickly using this algorithm. Notice that $\mathbf{X}_k$ is independent of β and so

$$\frac{\partial\Phi(\mathbf{C}^*_{d,\xi}(z), \beta)}{\partial\beta} = \mathrm{tr}\Big(\frac{\partial\tilde{\mathbf{H}}(\beta)}{\partial\beta}\mathbf{X}_k\tilde{\mathbf{H}}^T(\beta) + \tilde{\mathbf{H}}(\beta)\mathbf{X}_k\frac{\partial\tilde{\mathbf{H}}^T(\beta)}{\partial\beta} + \frac{\partial\tilde{\mathbf{J}}(\beta)}{\partial\beta}\tilde{\mathbf{J}}^T(\beta)$$
$$+\tilde{\mathbf{J}}(\beta)\frac{\partial\tilde{\mathbf{J}}^T(\beta)}{\partial\beta}\Big)$$

using the commutativity of the trace operator and the derivative operator.

Algorithm 4.1

```
Phase 1

FOR $\xi = 0$ TO $T - \xi_1$ STEP $\xi_1$ (where $\xi_1 \approx 0.05 * T$)
    CALCULATE $\mathbf{C}^*_{d,\xi}(z)$
    CALCULATE $\Phi_1(\mathbf{C}^*_{d,\xi}(z))$ for N$\approx 2$
NEXT $\xi$
RECORD BEST $\xi$, ($\overset{\triangle}{=} \xi^*$)

Phase 2

FOR $\xi = \max\{\xi^* - \xi_1, 0\}$ TO $\min\{\xi^* + \xi_1, T - \xi_2\}$ STEP $\xi_2$ (where $\xi_2 \approx 0.005T$)
        CALCULATE $\mathbf{C}^*_{d,\xi}(z)$
        CALCULATE $\Phi_1(\mathbf{C}^*_{d,\xi}(z))$ for $N \approx 10$
NEXT $\xi$
RECORD BEST $\xi$
```

Algorithm 4.2

```
FOR $\xi = 0$ TO T-$\xi_1$ STEP $\xi$ (where $\xi_1 \approx 0.05 * T$)
        FIND $\mathbf{C}^*_{d,\xi}(z)$
        $\beta := 0.5 * T$
        via the calculation of $\Phi(\mathbf{C}^*_{d,\xi}(z), \beta)$ and $\partial\Phi(\mathbf{C}^*_{d,\xi}(z), \beta)/\partial\beta$ perform
        a gradient ascent type algorithm to find $\Phi_2$
NEXT $\xi$
RECORD BEST $\xi$
```

The method outlined earlier for the computation of $\partial\Phi(C^*_{d,\xi_1}(z), \beta)/\partial\beta$ can be used in evaluating the supremum in $\Phi_2(\mathbf{C}^*_{d,\xi}(z))$.

Although both algorithms are basically linear search type algorithms, they are found to be rapidly convergent due to fact that the optimization is over only one variable ξ in a finite domain $\xi \in [0, T)$.

4.3 Convex Optimization Method

4.3.1 Introduction

The authors of [13] observe that there are two basic schemes for the design of linear time invariant controllers:

1. The first involves some form of parameter optimization. Whilst simple systems with only a small number of parameters can be effectively designed in this manner, more complicated multi-input, multi-output systems with many parameters pose problems. Even with the modern computer, systems which have a parameterized controller often have closed-loop cost functions which are non-convex in the controller parameters. Consequently, global minima are difficult to find.
2. The second form of control design is based on analytical methods which optimize with respect to some well defined criteria. These are generally quite limited in scope. It is often difficult to impose design constraints using such methods.

The claim of the work of [13] is that a wide variety of controller design problems can be formulated as a convex problem. The classical H_2, H_∞ and L_1 objectives form part of the available design objectives. Furthermore, design constraints such as asymptotic tracking, overshoot, undershoot, settling-time, signal peaks, e.t.c. can be incorporated. A disadvantage of this scheme is the fact that controllers of high order are generated but as the authors point out, this may not always be a problem. The resulting controller may be reduced in order through some form of model order reduction. Alternatively a special purpose architecture may be implemented.

This section demonstrates that the controller discretization problem can be formulated in terms of a convex optimization problem and can be solved efficiently using convex optimization algorithms. The motivation behind this work and the convex optimization algorithms used are found in [13, 14]. This work is based upon the parameterization of all controllers which stabilize a given plant [25, 83, 87, 88].

As mentioned in the introduction to this chapter, the numerical difficulties that are sometimes encountered in the previous method provide motivation for the use of a convex optimization approach to the problem of controller discretization. The optimal value of the primary cost function can be very sensitive to variations in the optimal

Youla parameter $\mathbf{Q}$. In some examples, the optimal controller is difficult to find. The algorithm to be presented does not suffer from these problems.

A further point is in relation to the unstable reference model case of the previous method. It is clear that the interpolation constraints may be difficult to enforce, particularly in the multivariable setting. However, it will be demonstrated that the convex optimization method automatically satisfies these constraints under certain conditions.

The convex optimization method can be formulated in an identical manner to the previous method. However a slightly more general formulation is presented which allows different selections of $\mathfrak{F}_c$.

4.3.2 Problem Formulation

Again define $\mathbf{P}^s_T(z)$ and $\mathbf{P}^s_{T,\xi}(z)$ to be the step invariant transformation and ξ-offset step invariant transformations of $\mathcal{P}(s)$ respectively (with sampling period T). Let $\mathbf{R}_T(z)$ be the impulse invariant transformation of $\mathcal{R}(s)$. Let the closed-loop transfer function $\mathcal{H}(s)$ of the analog system be defined by

$$\mathcal{H}(s) \triangleq (\mathbf{I}_n + \mathcal{P}(s)\mathcal{C}(s))^{-1}\mathcal{P}(s)\mathcal{C}(s) \tag{4.3.1}$$

and define

$$\mathcal{Y}(s) \triangleq \mathcal{H}(s)\mathcal{R}(s) \tag{4.3.2}$$

Finally, define

$$\mathbf{Y}_{T,\xi}(z) \triangleq Z_{T,\xi}\{\mathcal{Y}(s)\} \tag{4.3.3}$$

which can be written as

$$\mathbf{Y}_{T,\xi}(z) = \mathbf{H}^r_{T,\xi}(z)\mathbf{R}_T(z) \tag{4.3.4}$$

where $\mathbf{H}^r_{T,\xi}(z)$ is the ξ-offset signal invariant transformation of $\mathcal{H}(s)$ with respect to $\mathbf{r}(t)$.

With these definitions, the impulse response of $\mathbf{Y}_{T,\xi}(z)$ generates the sequence $y(kT+\xi)$; $k = 0, 1, \ldots, \infty$. Moreover, the impulse response of the system formed by

$$\mathbf{P}^s_{T,\xi}(z)(\mathbf{I}_n + \mathbf{C}_d(z)\mathbf{P}^s_T(z))^{-1}\mathbf{C}_d(z)\mathbf{R}_T(z) \tag{4.3.5}$$

generates the sequence $\hat{\mathbf{y}}(kT+\xi)$; $k = 0, 1, \ldots, \infty$.

Define an error system $\mathbf{E}_{T,\tilde{\xi}}(z)$ by

$$\mathbf{E}_{T,\tilde{\xi}}(z) \triangleq \left[\begin{array}{cccc} \mathbf{E}^T_{T,\xi_1}(z) & \mathbf{E}^T_{T,\xi_2}(z) & \cdots & \mathbf{E}^T_{T,\xi_N}(z) \end{array} \right]^T \tag{4.3.6}$$

where

$$\mathbf{E}_{T,\xi_i}(z) \triangleq \mathbf{P}^s_{T,\xi_i}(z)(\mathbf{I}_n + \mathbf{C}_d(z)\mathbf{P}^s_T(z))^{-1}\mathbf{C}_d(z)\mathbf{R}_T(z) - \mathbf{Y}_{T,\xi_i}(z) \; ; \; i = 1, 2, \cdots, N \tag{4.3.7}$$

Figure 4-2, shows a diagrammatic representation of $\mathbf{E}_{T,\tilde{\xi}}(z)$. In this diagram, the z_i for $i = 1 \ldots N$ are the difference in the outputs of the analog system and the hybrid system for a series of intersample points $\xi_1, \xi_2, \ldots, \xi_N$, $0 \leq \xi_i < T$. The selection of the ξ_i encompasses some freedom, but generally a uniform distribution of ξ_i in the interval $[0, T)$ is adequate. Typically $N \approx 4-6$ is adequate. The input $w(kT)$ is the Kronecker delta function.

If $\tilde{z}$ is defined as

$$\tilde{\mathbf{z}} \triangleq \left[\begin{array}{cccc} \mathbf{z}_1 & \mathbf{z}_2 & \ldots & \mathbf{z}_N \end{array} \right]^T \tag{4.3.8}$$

then $\mathbf{E}_{T,\tilde{\xi}}(z)$ is the transfer function matrix from w to $\tilde{\mathbf{z}}$. The problem of this section is now stated.

Find a stabilizing digital control $\mathbf{C}_d(z)$ which minimizes the norm of $\mathbf{E}_{T,\tilde{\xi}}(z)$, where $\mathbf{E}_{T,\tilde{\xi}}(z)$ is the transfer matrix from w to $\tilde{\mathbf{z}}$ in Figure 4-2.

As already stressed, a major advantage of the convex optimization approach is that there is great freedom in the choice of norm. Again there are some restrictions and considerations in tracking unstable reference signals—a short discussion of these is given in Section 4.3.4. Recall that constraints may also be imposed during the minimization.

Figure 4-2 can be transformed into Figure 4-3, a standard two-port system. The quantity $\tilde{\mathbf{P}}(z)$ is a generalized plant with description

$$\left[\begin{array}{c} \tilde{\mathbf{z}} \\ \mathbf{y} \end{array} \right] = \tilde{\mathbf{P}}(z) \left[\begin{array}{c} w \\ \mathbf{u} \end{array} \right] = \left[\begin{array}{cc} \tilde{\mathbf{P}}_{\tilde{z}w} & \tilde{\mathbf{P}}_{\tilde{z}u} \\ \tilde{\mathbf{P}}_{yw} & \tilde{\mathbf{P}}_{yu} \end{array} \right] \left[\begin{array}{c} w \\ \mathbf{u} \end{array} \right] \tag{4.3.9}$$

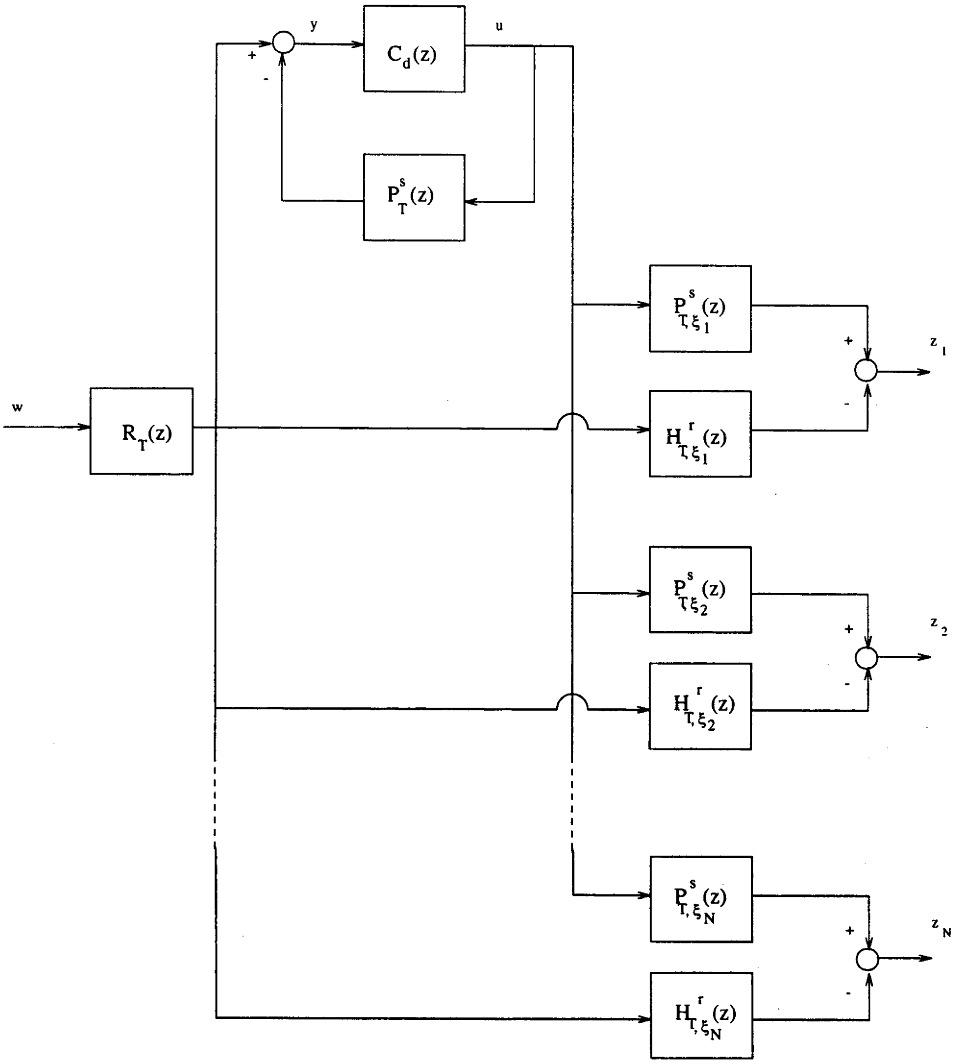

Figure 4-2: Analog and digital control

and $\tilde{\mathbf{P}}(z)$ is given by

$$\tilde{\mathbf{P}}(z) \triangleq \begin{bmatrix} \begin{bmatrix} -\mathbf{H}^r_{T,\xi_1}(z) & \mathbf{P}^s_{T,\xi_1}(z) \\ -\mathbf{H}^r_{T,\xi_2}(z) & \mathbf{P}^s_{T,\xi_2}(z) \\ \cdot & \cdot \\ \cdot & \cdot \\ -\mathbf{H}^r_{T,\xi_N}(z) & \mathbf{P}^s_{T,\xi_N}(z) \end{bmatrix} \times \mathbf{R}_T(z) \\ \mathbf{R}_T(z) \quad -\mathbf{P}^s_T(z) \end{bmatrix} \tag{4.3.10}$$

Let $\tilde{\mathbf{P}}(z)$ have a state space realization given by

$$\mathbf{x}(kT+T) = \mathbf{A}\mathbf{x}(kT) + \mathbf{B}_w w(kT) + \mathbf{B}_u \mathbf{u}(kT) \tag{4.3.11}$$

$$\tilde{\mathbf{z}}(kT) = \mathbf{C}_{\tilde{z}}\mathbf{x}(kT) + \mathbf{D}_{\tilde{z}w} w(kT) + \mathbf{D}_{\tilde{z}u} u(kT) \tag{4.3.12}$$

$$\mathbf{y}(kT) = \mathbf{C}_y\mathbf{x}(kT) + \mathbf{D}_{yw} w(kT) + \mathbf{D}_{yu} \mathbf{u}(kT) \tag{4.3.13}$$

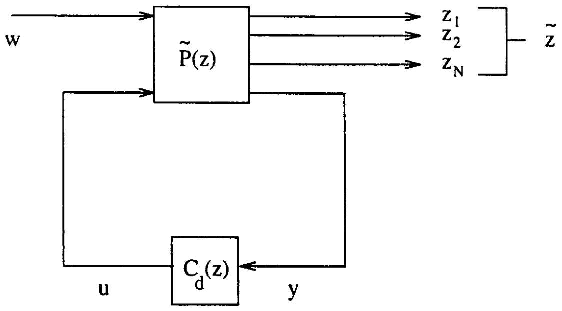

Figure 4-3: Analog and digital control

Let the order of $\mathcal{H}(s), \mathcal{R}(s), \mathcal{P}(s)$ and $\tilde{\mathbf{P}}(z)$ be n_h, n_r, n_p and $n_{\tilde{p}}$ respectively. Then

$$n_{\tilde{p}} \leq n_h + n_r + n_p$$

As evidenced from Figure 4-2, $\tilde{\mathbf{P}}(z)$ contains $n_r + n_h$ modes that are uncontrollable from $\mathbf{u}$—the states of $\mathbf{R}_T(z)$ and the states of $\mathbf{H}^r_{T,\xi_i}(z)$, $i = 1, 2, \cdots, N$. Following from assumption A in Section 2.3 of Chapter 2, the states of $\mathbf{H}^r_{T,\xi_i}(z)$ are stabilizable. If $\mathcal{R}(s)$ is stable then $\tilde{\mathbf{P}}(z)$ is stabilizable. Again the discussion of unstable $\mathcal{R}(s)$ is postponed until Section 4.3.4.

4.3.3 The Use of "qdes"

The authors of [13] have provided a convex optimization package "qdes" which is available at the public ftp site "isl.stanford.edu". In the directory "/pub/boyd/qdes" is the series of programs compressed in "qdes_dist.tar.Z". In order to use this material, the following results are needed.

Assume $\mathbf{K}$ is chosen such that $\mathbf{A} - \mathbf{B}_u\mathbf{K}$ is stable and $\mathbf{L}$ such that $\mathbf{A} - \mathbf{L}\mathbf{C}_y$ is stable. Then from [83] a doubly coprime factorization of $\tilde{\mathbf{P}}_{yu}$ can be found such that

$$\tilde{\mathbf{P}}_{yu} = \mathbf{N}\mathbf{D}^{-1} = \tilde{\mathbf{D}}^{-1}\tilde{\mathbf{N}} \tag{4.3.14}$$

with

$$\begin{bmatrix} \mathbf{Y} & \mathbf{X} \\ -\tilde{\mathbf{N}} & -\tilde{\mathbf{D}} \end{bmatrix} \begin{bmatrix} \mathbf{D} & -\tilde{\mathbf{X}} \\ \mathbf{N} & \tilde{\mathbf{Y}} \end{bmatrix} = \mathbf{I}_{m+n} \tag{4.3.15}$$

where

$$\begin{aligned}
\tilde{\mathbf{N}}(z) &= \mathbf{C}_y(z\mathbf{I}_{n_{\bar{p}}} - \mathbf{A} + \mathbf{L}\mathbf{C}_y)^{-1}(\mathbf{B}_u - \mathbf{L}\mathbf{D}_{yu}) + \mathbf{D}_{yu} && (4.3.16)\\
\tilde{\mathbf{D}}(z) &= \mathbf{I}_n - \mathbf{C}_y(z\mathbf{I}_{n_{\bar{p}}} - \mathbf{A} + \mathbf{L}\mathbf{C}_y)^{-1}\mathbf{L} && (4.3.17)\\
\mathbf{N}(z) &= (\mathbf{C} - \mathbf{D}_{yu}\mathbf{K})(z\mathbf{I}_{n_{\bar{p}}} - \mathbf{A} + \mathbf{B}_u\mathbf{K})^{-1}\mathbf{B}_u + \mathbf{D}_{yu} && (4.3.18)\\
\mathbf{D}(z) &= \mathbf{I}_m - \mathbf{K}(z\mathbf{I}_{n_{\bar{p}}} - \mathbf{A} + \mathbf{B}_u\mathbf{K})^{-1}\mathbf{B}_u && (4.3.19)\\
\mathbf{X}(z) &= \mathbf{K}(z\mathbf{I}_{n_{\bar{p}}} - \mathbf{A} + \mathbf{L}\mathbf{C}_y)^{-1}\mathbf{L} && (4.3.20)\\
\mathbf{Y}(z) &= \mathbf{I}_m + \mathbf{K}(z\mathbf{I}_{n_{\bar{p}}} - \mathbf{A} + \mathbf{L}\mathbf{C}_y)^{-1}(\mathbf{B}_u - \mathbf{L}\mathbf{D}_{yu}) && (4.3.21)\\
\tilde{\mathbf{X}}(z) &= \mathbf{K}(z\mathbf{I}_{n_{\bar{p}}} - \mathbf{A} + \mathbf{B}_u\mathbf{K})^{-1}\mathbf{L} && (4.3.22)\\
\tilde{\mathbf{Y}}(z) &= \mathbf{I}_n + (\mathbf{C}_y - \mathbf{D}_{yu}\mathbf{K})(z\mathbf{I}_{n_{\bar{p}}} - \mathbf{A} + \mathbf{B}_u\mathbf{K})^{-1}\mathbf{L} && (4.3.23)
\end{aligned}$$

From [2], a controller which internally stabilizes $\tilde{\mathbf{P}}_{yu}$ is such that

$$\begin{bmatrix} \mathbf{I}_m & \mathbf{C}_d(z) \\ -\tilde{\mathbf{P}}_{yu}(z) & \mathbf{I}_n \end{bmatrix}^{-1} \qquad (4.3.24)$$

is stable and proper. One such controller is given by

$$\begin{aligned}
\mathbf{x}_c(kT+T) &= (\mathbf{A} - \mathbf{B}_u\mathbf{K} - \mathbf{L}\mathbf{C}_y)\mathbf{x}_c(kT) + \mathbf{L}\mathbf{y}(kT) && (4.3.25)\\
\mathbf{u}(kT) &= \mathbf{K}\mathbf{x}_c(kT) && (4.3.26)
\end{aligned}$$

From [83], the class of all stabilizing controllers is given by

$$\mathbf{C}_d(z) = (\mathbf{Y}(z) - \mathbf{Q}(z)\tilde{\mathbf{N}}(z))^{-1}(\mathbf{X}(z) + \mathbf{Q}(z)\tilde{\mathbf{D}}(z)) \qquad (4.3.27)$$

and the stable closed-loop transfer function from w to $\mathbf{z}$, $\mathbf{E}_{T,\tilde{\xi}}(z)$ is given by

$$\mathbf{E}_{T,\tilde{\xi}}(z) = \mathbf{T}_1(z) + \mathbf{T}_2(z)\mathbf{Q}(z)\mathbf{T}_3(z)\ ;\ \mathbf{Q}(z) \in RH^{\infty} \qquad (4.3.28)$$

with

$$\begin{bmatrix} \mathbf{T}_1(z) & \mathbf{T}_2(z) \\ \mathbf{T}_3(z) & \mathbf{0} \end{bmatrix} \triangleq \mathbf{C}_T(z\mathbf{I}_{2n_{\bar{p}}} - \mathbf{A}_T)^{-1}\mathbf{B}_T + \mathbf{D}_T \qquad (4.3.29)$$

where

$$\mathbf{A}_T = \begin{bmatrix} \mathbf{A} & -\mathbf{B}_u\mathbf{K} \\ \mathbf{L}\mathbf{C}_y & \mathbf{A}-\mathbf{B}_u\mathbf{K}-\mathbf{L}\mathbf{C}_y \end{bmatrix} \tag{4.3.30}$$

$$\mathbf{B}_T = \begin{bmatrix} \mathbf{B}_w & \mathbf{B}_u \\ \mathbf{L}\mathbf{D}_{yw} & \mathbf{B}_u \end{bmatrix} \tag{4.3.31}$$

$$\mathbf{C}_T = \begin{bmatrix} \mathbf{C}_{\tilde{z}} & \mathbf{D}_{zu}\mathbf{K} \\ \mathbf{C}_y & -\mathbf{C}_y \end{bmatrix} \tag{4.3.32}$$

$$\mathbf{D}_T = \begin{bmatrix} \mathbf{D}_{\tilde{z}w} & \mathbf{D}_{\tilde{z}u} \\ \mathbf{D}_{yw} & \mathbf{0} \end{bmatrix} \tag{4.3.33}$$

$$\tag{4.3.34}$$

The closed-loop transfer function is in a form that is affine in $\mathbf{Q}(z)$ (c.f. (4.3.28)). Hence an infinite dimensional convex optimization problem can be formulated to minimize (4.3.28) for some norm.

The authors of [14] give an outline of the Ritz method [23, 55] for obtaining an approximate solution to this problem. The basic idea is to approximate the infinite dimensional problem by a finite dimensional one, by restricting $\mathbf{Q}$ to a large, but finite dimensional, space. This is a achieved by setting

$$\mathbf{Q}(z) = \mathbf{Q}_0(z) + \sum_{i=1}^{n\_tap} x_i\mathbf{Q}_i(z) \; ; \; x_i \in \mathbb{R}, \; \mathbf{Q}_i(z) \in RH^{\infty} \tag{4.3.35}$$

The $\mathbf{Q}_i(z)$ are fixed basis functions and *n_tap* is the number of basis functions used. The problem is to find

$$\mathbf{x} \triangleq [x_1, x_2, \cdots, x_{n\_tap}]$$

the decision variable, which minimizes (4.3.28) subject to any constraints that may be imposed. Corresponding to the optimal $\mathbf{x}$ is the optimal controller $\mathbf{C}_{\mathrm{d}}^*(z)$. The numerical procedure outlined in [13] requires the evaluation of the impulse responses of $\mathbf{T}_1(z), \mathbf{T}_2(z)$ and $\mathbf{T}_3(z)$ to a length of *n_sample*. For further details refer to [13].

4.3.4 Unstable $\mathcal{R}(s)$

The theory presented to this point is suitable when the reference model $\mathcal{R}(s)$ is stable. To conclude this section on the convex optimization method of controller discretization, a short discussion of the unstable reference signal case is given.

As stated in Section 4.2.3, the unstable reference signal case can be dealt with by imposing interpolation constraints on the Youla parameter $\mathbf{Q}(z)$ or by ensuring that a continuous-time internal model of the reference signal is present in the open-loop dynamics.

If $N > 1$, then in general, an internal model must be present to ensure that $z_i \to 0$ ($i = 1, 2, \ldots, N$) as $t \to \infty$ (which gives asymptotic tracking). This is necessary for a finite $\|\mathbf{E}_{T,\xi}\|$ in the case of l_1 and l_2 norms. However there will be cases, especially for small N, in which interpolation constraints can be imposed to ensure a finite norm. The beauty of the Ritz method is that the interpolation conditions appear as linear constraints on the x_i. Once the $\mathbf{Q}_i(z)$ are specified, the poles of $\mathbf{Q}(z)$ are determined. The x_i then determine the zeros of $\mathbf{Q}(z)$. Provided *n_tap* is large enough, the Ritz method automatically enforces the interpolation constraints because the optimal solution must be contained in the subspace generated by the constraints. Thus, the convex optimization method of controller discretization is a very appealing algorithm from a numerical perspective.

4.4 Integral Approximation Method

4.4.1 Introduction

The underlying motivation of the method presented in this section is that the techniques of Sections 4.2 and 4.3 are in one sense indirect. That is, the optimization procedures produce digital controllers of high order which generally require some form of model order reduction. The techniques developed in this section approach the problem more directly. That is, the digital controller produced is of low order (typically equal to that of the analog controller), and as a result, no reduction of order is necessary in the final phase of the algorithm.

A secondary motivation is to bring the insights gained in the open-loop discretization method into the closed-loop setting. Therefore, the method developed in this section is primarily an extension of the theory developed in Chapter 3. Each integrator of the continuous-time controller is replaced by a modified Newton-Cotes type approximation. The order of the approximation of each integrator can vary, but is typically zero, one, or two. The modified Newton-Cotes formulae are parameterized and then optimized with respect to a given signal, generated via the impulse response of a linear, time-invariant system. Again the allocation of different discretization order to each integral is discussed, as well as the motivation for selecting certain state space structures before discretization.

4.4.2 Problem Formulation

Assume that the digital controller is in fact parameterized according to the method of Chapter 3 and represent the parameterized controller as $\mathbf{C}_d(\mathbf{p}, z)$, where $\mathbf{p}$ is the parameter vector.

The closed-loop operator $\mathfrak{F}_c$ is chosen to be a cost function $\mathcal{J}_N(\mathbf{C}_d(\mathbf{p}, z))$ where

$$\mathcal{J}_N(\mathbf{C}_d(\mathbf{p}, z)) = \max_i \sum_{k=0}^{N} \|\hat{\mathbf{y}}_p(kT+\xi_i) - \mathbf{y}(kT+\xi_i)\|_2^2 \quad 0 \le \xi_i < T,\ i = 1, 2, \ldots, M \tag{4.4.1}$$

with $\mathbf{y}(kT+\xi_i)$ the sampled output of the analog system at an intersample point ξ_i and $\hat{\mathbf{y}}_p(kT+\xi_i)$ the sampled output of the sampled-data system at an intersample point ξ_i.

The problem can be stated as follows:

For a given N and set of ξ_i, $i = 1, 2, \ldots, M$, find the parameter $\mathbf{p}$ and resulting $\mathbf{C}_d(\mathbf{p}, z)$ such that $\mathcal{J}_N(\mathbf{C}_d(\mathbf{p}, z))$ given by (4.4.1) is minimized.

4.4.3 The Algorithm

Given a parallel problem formulation to the one in Section 4.2.2, let the state space realization of

$$\mathbf{P}^s_{T,\xi_i}(z)(\mathbf{I} + \mathbf{P}^s_T(z)\mathbf{C}_d(\mathbf{p}, z))^{-1}\mathbf{C}_d(\mathbf{p}, z)\mathbf{R}_T(z) - \mathbf{Y}_{T,\xi_i}(z)\ ;\ i = 1, 2, \ldots, M$$

be $[\mathbf{F}_i, \mathbf{G}_i, \mathbf{H}_i, \mathbf{J}_i]$. The computation of (4.4.1) involves the computation of the sum of impulse responses squared for each of the M systems $[\mathbf{F}_i, \mathbf{G}_i, \mathbf{H}_i, \mathbf{J}_i]$.

As is shown in Chapter 3, this can be effectively calculated using a doubling algorithm for Lyapunov equations. Again, gradient calculation can be readily extracted. The doubling algorithm is effective for large N with $N \approx 2^{10-15}$ giving a good approximation to the H_2 norm. Because N is restricted to be finite, unstable $\mathcal{R}(s)$ poses few problems compared to other discretization methods. Conversely, there is no guarantee of finding a stabilizing controller. However if N is chosen large, a stabilizing controller is generally found.

The **minimax** algorithm of MATLAB has been found to be quite effective in the solution from this point. The problem of local minima is more of a problem here than in the open-loop problem. However a simple technique has been found which greatly alleviates this problem.

The technique is to run the minimax algorithm in two passes. The first pass is done with a small value of k, typically 2-3. The final parameter value of the first pass is then used as the initial parameter for a larger value of k. It appears as if the result of the first pass places the initial parameter to the second pass in the neighbourhood of the basin of attraction of the global minimum for large k. The second pass then rapidly finds the global minimum.

A value of $M \approx 4$ seems to work well for most problems, with the intersample points ξ_i uniformly spaced in the interval $0 \leq \xi_i < T$.

In Chapter 3, a number of factors which affect discretization error were identified—these results are applicable to the open-loop discretization problem. These factors enable the designer to effectively allocate different order integral approximations to the different states. For the closed-loop problem the results are not so conclusive, although it generally appears as if the Hankel singular values weighted by the closed-loop transfer function and the controller state space structure play the greatest roles. The same criterion (weighted Hankel singular values) used for the controller reduction schemes [3, 27] are generally a good guide for selecting which order discretization should be applied to which states. Furthermore, simulation studies reveal that, again, the internally balanced state space structure is generally a good one.

4.5 Simulation Studies

This section presents the results of simulation studies that were obtained using the controller discretization methods developed in this chapter. The simulations were performed using MATLAB (a description of the MATLAB code can be found in Appendix F). The methods were applied to the benchmark problems of Rattan [65] and Katz [42].

These studies are only a preliminary investigation into the performance of the algorithms. A more complete study is undertaken in Chapter 5.

The Rattan example is a unity feedback control system with plant's transfer function given by

$$\mathcal{H}(s) = \frac{10}{s(s+1)} \tag{4.5.1}$$

and the analog controller's transfer function given by

$$\mathcal{C}(s) = \frac{0.416s+1}{0.139s+1} \tag{4.5.2}$$

The Katz example has the same unity feedback configuration with

$$\mathcal{H}(s) = \frac{863.3}{s^2} \tag{4.5.3}$$

and controller

$$\mathcal{C}(s) = \frac{2940s+86436}{(s+294)^2} \tag{4.5.4}$$

4.5.1 H_2 Optimal Control Method

For both benchmark examples, digital controllers were designed using Algorithm 4.1. Simulation results are compared to to three existing methods, [1, 45, 65].

Rattan's Example

As with the original paper [65], the digital controller was designed with a sampling rate of $T = 0.157s$. The optimization was performed with respect to a step, i.e. $\mathcal{R}(s) = 1/s$. The resulting third order digital controller is

$$C_d(z) = \frac{2.273z^3 - 1.826z^2 - 1.249z + 0.982}{z^3 - 0.027z^2 - 0.785z + 0.135} \tag{4.5.5}$$

In Figure 4-4 a comparison between the new method and the methods of Anderson and Keller, and of Rattan is shown. It should be noted that the Anderson and Keller digital controller (designed via the "stability margin method") is also third order while the Rattan digital controller is first order.

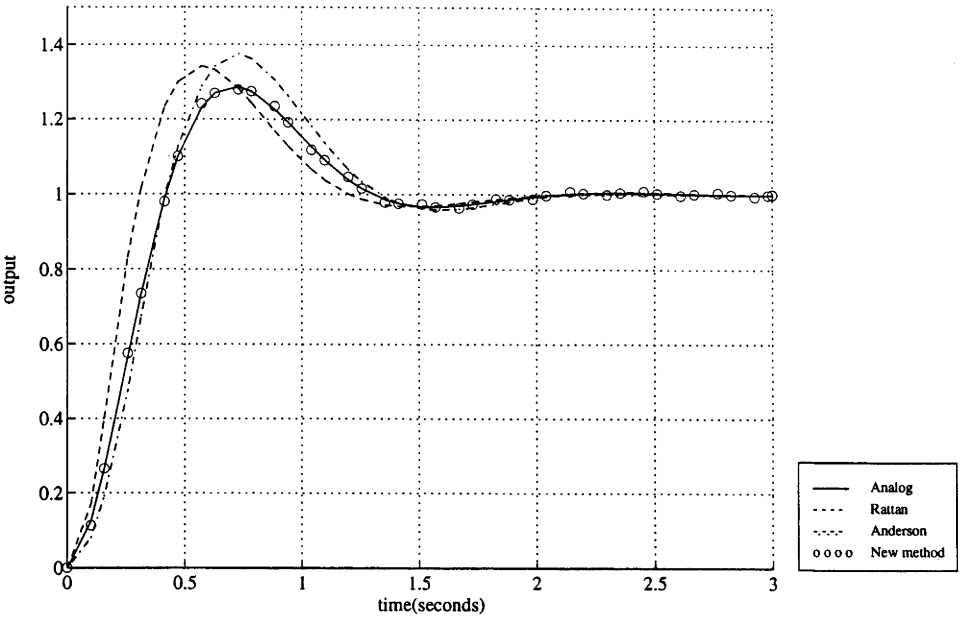

Figure 4-4: Rattan's example - Step responses of existing discretization schemes compared to the H_2 optimal control method, $T = 0.157$s

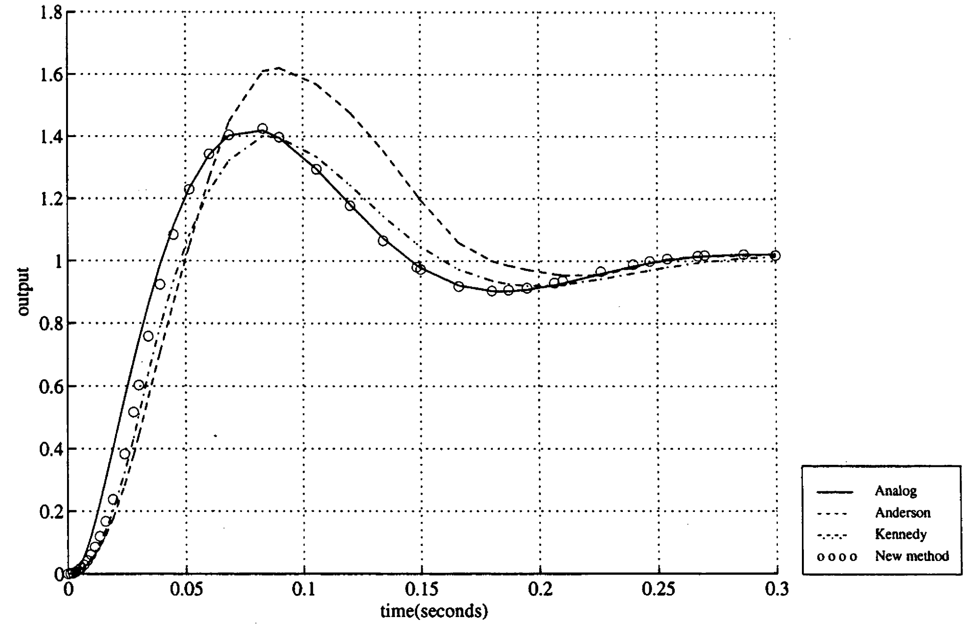

Figure 4-5: Katz example - Step responses of existing discretization schemes compared to the H_2 optimal control method, $T = 0.03$s

Katz Example

Initially a digital controller was designed with a sampling rate of $T = 0.03s$. The resulting digital controller is

$$C_d(z) = \frac{1.551z^2 - 2.160z + 0.609}{z^2 - 0.225z - 0.306} \tag{4.5.6}$$

In Figure 4-5 a comparison between the new method and the existing methods is shown. The corresponding control signals for the analog and digital controllers are shown In Figure 4-6.

Figure 4-7 shows simulation results for a digital controller designed with the sampling period increased to $T = 0.05s$. At this sampling period, the other schemes can not provide a stabilizing controller. Results for a third order digital controller and a second order controller corresponding to

$$C_d(z) = \frac{0.741z^3 + 0.126z^2 - 0.701z - 0.166}{z^3 + 1.694z^2 + 0.793z + 0.0794} \tag{4.5.7}$$

and

$$C_d(z) = \frac{0.734z^2 - 0.090z - 0.643}{z^2 + 1.434z + 0.478} \tag{4.5.8}$$

respectively, are plotted. The degradation of performance, as the controller order is decreased, is evident.

A frequency-domain comparison is also made for this example. Obviously a problem arises between the comparison of a continuous-time system with a sampled-data system. For a discussion of the frequency response methods used in these studies refer to Appendix C. In Figure 4-8 the frequency responses of the following quantities are plotted:

1. the continuous-time closed-loop system;
2. the step invariant transformation of the continuous-time closed-loop system;
3. closed-loop system formed with the digital controller designed using existing discretization schemes, and the continuous-time plant replaced with its step invariant transformation equivalent.

It can be seen that the digital controller generated with the H_2 optimal control method (with $T = 0.03s$) has a fairly similar frequency response to Kennedy's two degree of

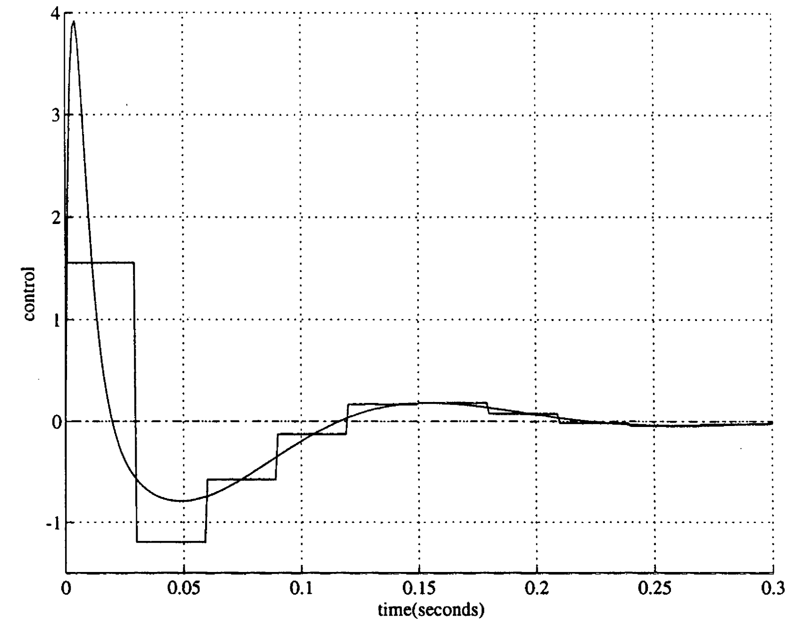

Figure 4-6: Katz example - Analog control signal and digital control signal of H_2 optimal control method

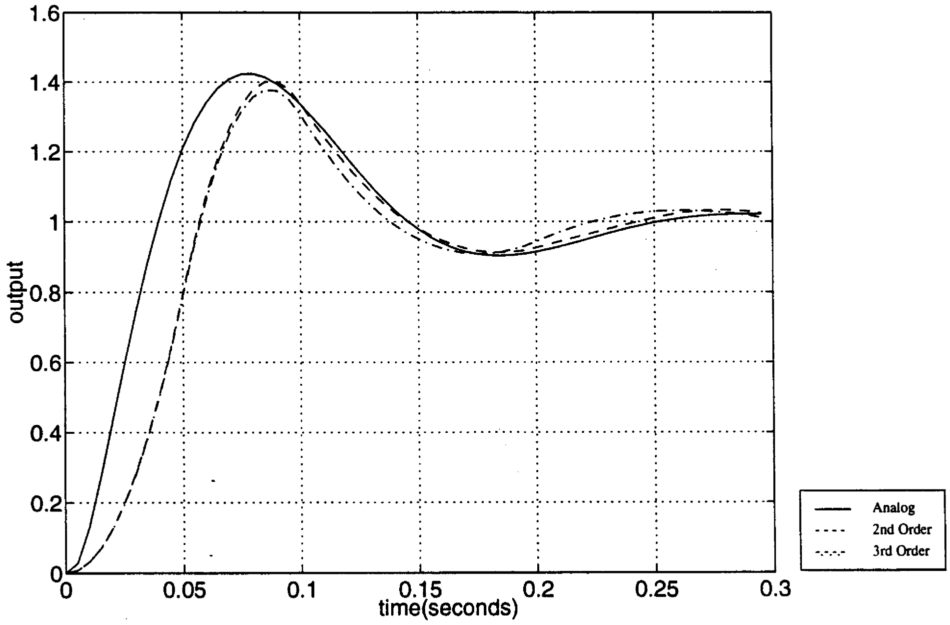

Figure 4-7: Katz example - Step responses of H_2 optimal control method with T = 0.05s

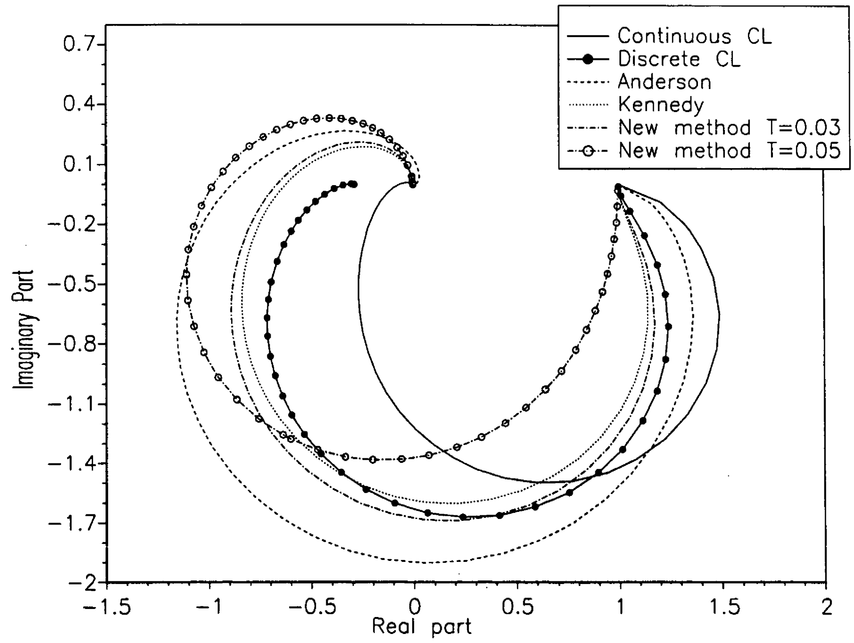

Figure 4-8: Katz example - Closed-loop frequency responses of existing discretization schemes and H_2 optimal control method

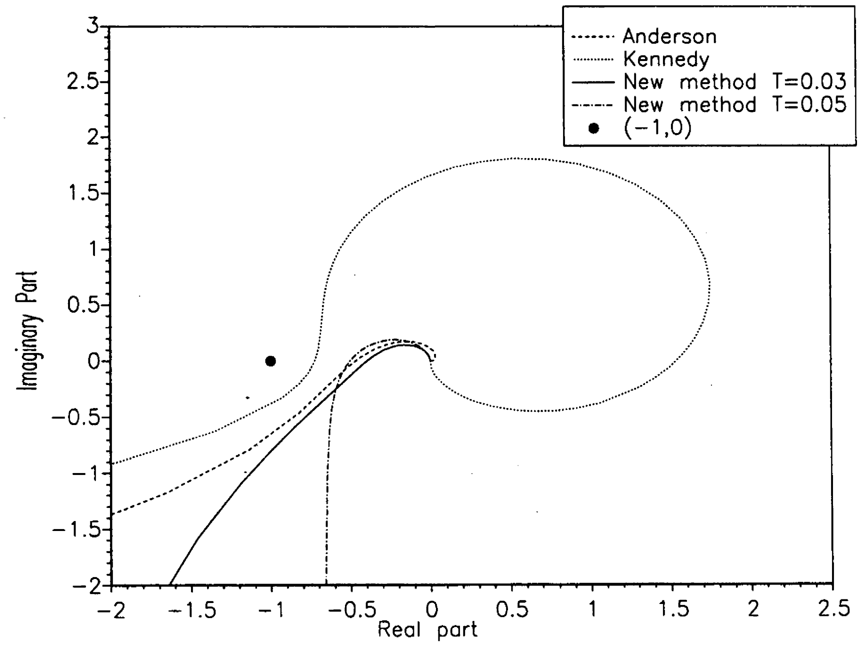

Figure 4-9: Katz example - Nyquist plots of existing discretization schemes and H_2 optimal control method

freedom controller. The fact that the controller designed with $T = 0.05s$ has a fairly acceptable performance is significant.

Also shown in Figure 4-9 is the Nyquist plots of the discrete-time loop transfer functions. Of particular importance is that digital controller (with $T = 0.03s$) designed using the H_2 optimal control method has superior gain and phase margins compared to other schemes. Also with $T = 0.05s$ the design has better margins than those of the Kennedy design with $T = 0.03s$ and better phase margin than that of Anderson with $T = 0.03s$.

Even though the design method is based upon time-domain methods, it has been shown that the frequency-domain results are also good. This is accounted for in part by the fact that an l_2 time-domain design quite often has good frequency-domain properties (c.f. Parseval's Theorem).

4.5.2 Convex Optimization Method

Using the "qdes" package in conjunction with the theory presented in Section 4.3.2, a practically very satisfactory method of controller discretization is achieved. As already stated, a variety of optimization problems can be formulated to yield a discretized controller. In this section the convex optimization procedures are used for the controller discretization problem of Katz.

Again the design was initially performed with a sampling rate of $T = 0.03s$. The "qdes" program enables a number of norms to be chosen for the minimization of $||E_{T,\tilde{\xi}}(z)||$. The H_2 norm was chosen, as the Katz problem allows this (so the objective cost is simply the sum of squares of the N regulated outputs z_i). Initially no constraints were imposed. The selection $N = 4$ was adequate for this problem. An initial controller was generated using equations (4.3.25)–(4.3.26). The convex optimization parameters *n_tap* and *n_sample* were set at

$$n\_tap = 20 \; ; \; n\_sample = 300$$

Making the the number of taps of Q larger resulted in little improvement for this example.

The convex optimization algorithm was executed and a stable digital controller of order equal to 27 was generated. The total computation time was 6.72 seconds CPU on a SPARC 10. In Figure 4-10 the step response is shown and in Figure 4-13 the corresponding digital control signal. Figure 4-12 show the step response of the digital controller designed with a sampling period increased to $T = 0.05$ seconds. As seen, the convex optimization method yields a very acceptable sampled-data system performance.

Model Order Reduction

As previously mentioned, the sampled-data model order reduction method [56] proves very effective for the purposes of this thesis. A frequency weighted balanced truncation technique was applied to the stable part of $C(z)$ (in fact the 27^{th} order controller generated in the last two examples is stable). The method was applied to the controller corresponding to $T = 0.03$ seconds. A second order controller $C_r(z)$ such that

$$||W(z)[C(z) - C_r(z)]V(z)||_\infty$$

is minimized was found, where $V(z), W(z)$ are weighting function generated according to the algorithm. The fast sampling factor N (a parameter associated with the model order reduction algorithm), was set equal 6. The resulting digital controller is given by

$$C_r(z) = \frac{1.7343z^2 - 2.9997z + 0.5778}{z^2 - 0.0468z - 0.4327}$$

The step response and control signals are plotted in Figures 4-13 and 4-14 respectively.

Constraint introduction

Assume the digital control signal is to be limited. This is easily incorporated in by adding an extra regulated output z_{N+1} corresponding to the control signal u. In this example a constraint of

$$|u| \leq 1$$

was enforced during the design(with a sampling period of $T = 0.03s$). Figures 4-15 and 4-16 show the resulting step response and control signal respectively.

4.5.3 Integral Approximation Method

Finally the integral approximation method was applied to the Katz example. Again a sampling rate of $T = 0.03s$ was chosen. A selection of $\bar{\mathbf{A}} = \mathbf{A}$ was chosen and all state

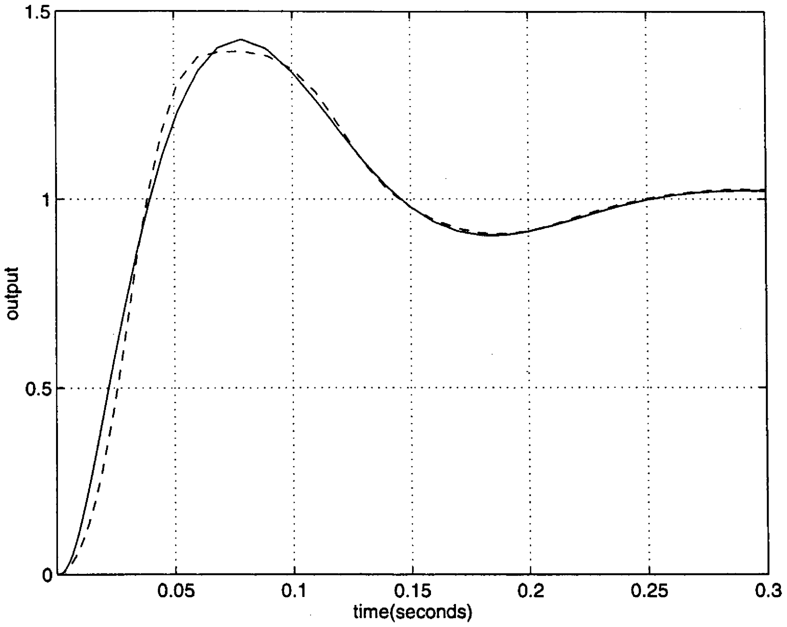

Figure 4-10: Katz example - Step response of convex optimization method, 27^{th} order controller, $T = 0.03$s

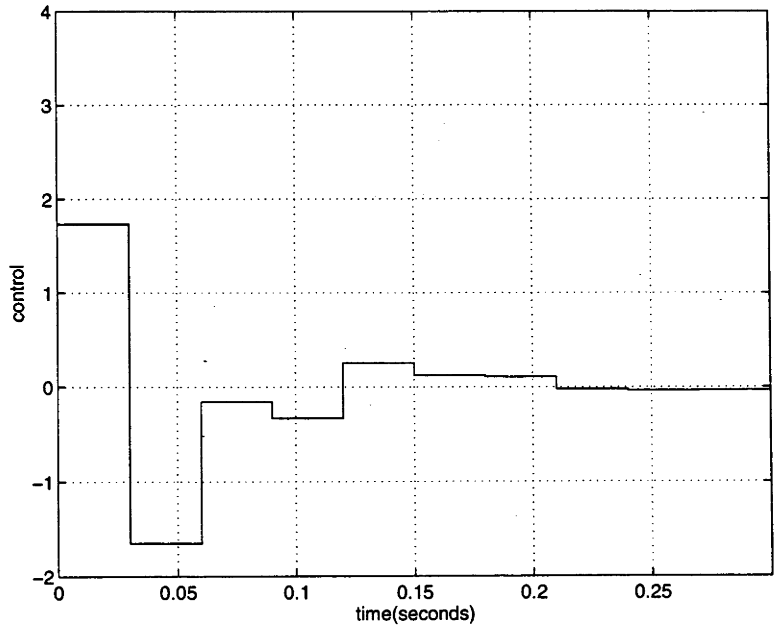

Figure 4-11: Katz example - Digital control signal of convex optimization method

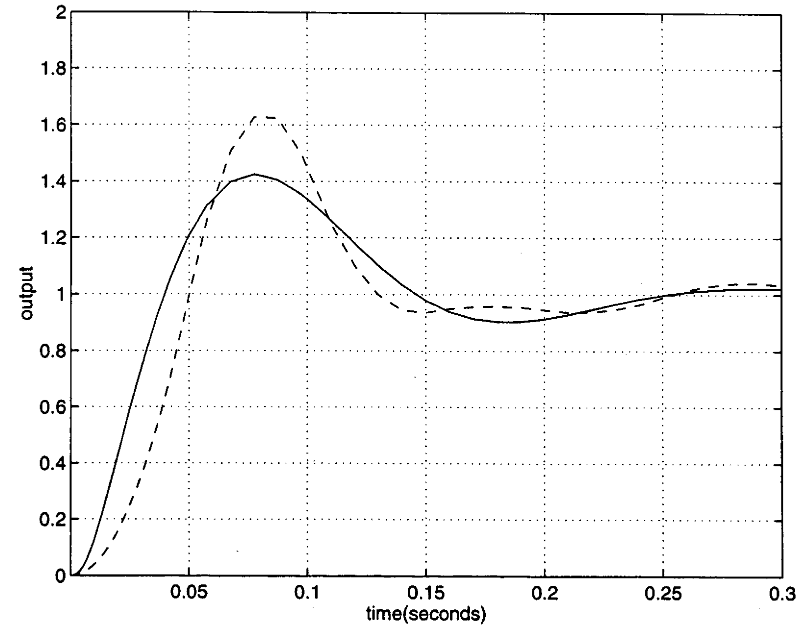

Figure 4-12: Katz example - Step response of convex optimization method with $T = 0.05s$

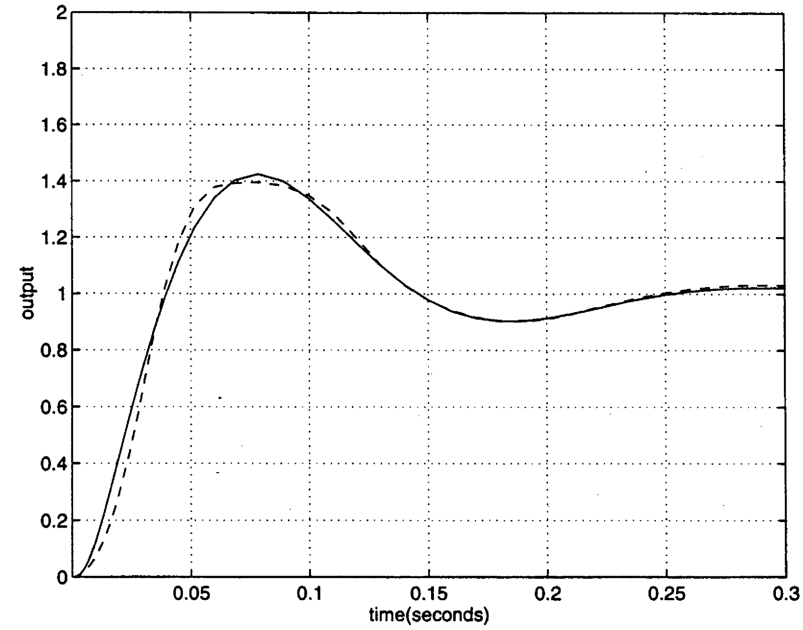

Figure 4-13: Katz example - Step response of convex optimization method, model order reduction applied to controller to bring it down to 2^{nd} order, $T = 0.03s$

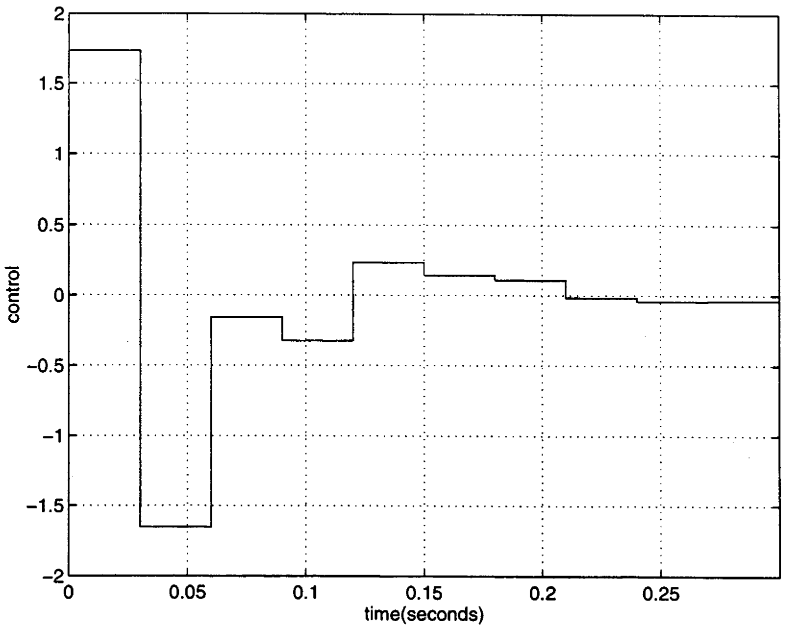

Figure 4-14: Katz example - Control signal of model order reduced controller

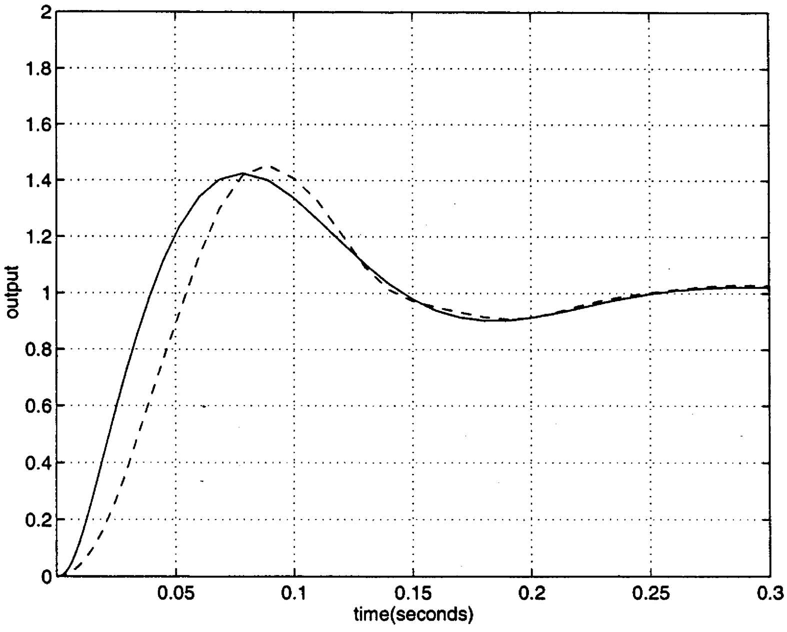

Figure 4-15: Katz example - Step response for convex optimization method with limited control signal, $|u| < 1$, $T = 0.03$s

integrals were approximated by first order. Selections of $M = 4$ and $N = 2^9 - 1$ were used. The resulting digital controller is

$$C_d(z) = \frac{1.2757 \times 10^0 z^2 - 1.2299 \times 10^0 - 2.4593 \times 10^{-5}}{z^2 - 2.9550 \times 10^{-4} + 2.1830 \times 10^{-8}} \tag{4.5.9}$$

In Figure 4-17 the step responses of the analog system and the hybrid system are shown. This result compares favourably to the other methods developed in this thesis and especially with existing methods published in the literature.

4.6 Conclusions

In this chapter, three methods of controller discretization have been presented. The methods are such that the closed-loop properties of the original system are respected. Each method is based on optimization principles and the optimization is primarily in the time-domain, although the convex optimization scheme does allow frequency-domain criteria.

Experience indicates that, from the point of view of tracking a reference signal, the proposed methods outperform other methods found in the literature. The signal invariant transformation method offers flexibility and good performance. The price to be paid is that the controller may be of high order. Model order reduction, taking into account the closed-loop properties to be preserved, can alleviate this problem. The frequency-domain properties of the design using this method are shown to be good as well.

The efficacy of the convex optimization method for the digital re-design of analog controllers has been illustrated. The method is numerically reliable and fast, offers great flexibility, and when incorporated with sampled-data model order reduction techniques, offers a very practical method of digital controller design.

The integral replacement method has the benefit that the resulting digital controller is of low order and so avoids the need of model order reduction in completing the design. The digital controller is shown to possess good performance on the benchmark example of Katz.

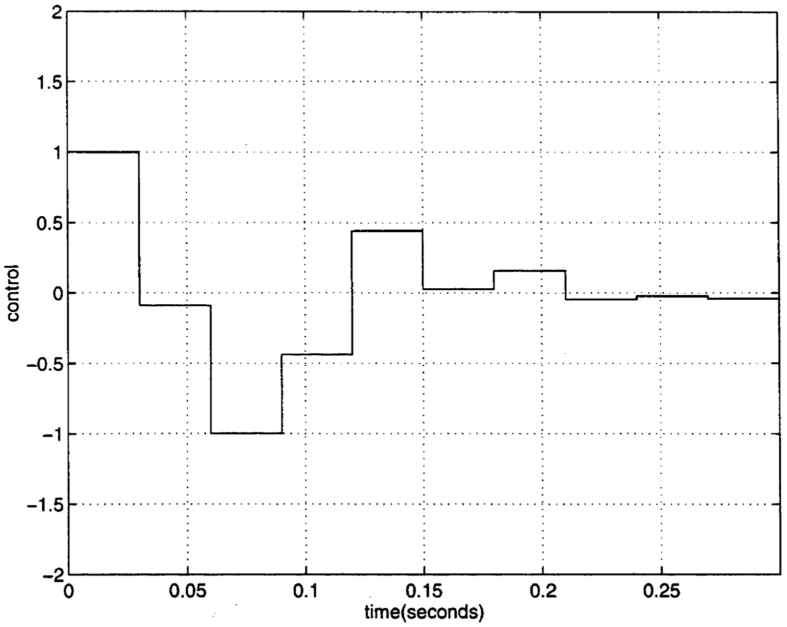

Figure 4-16: Katz example - Convex optimization method with limited control signal

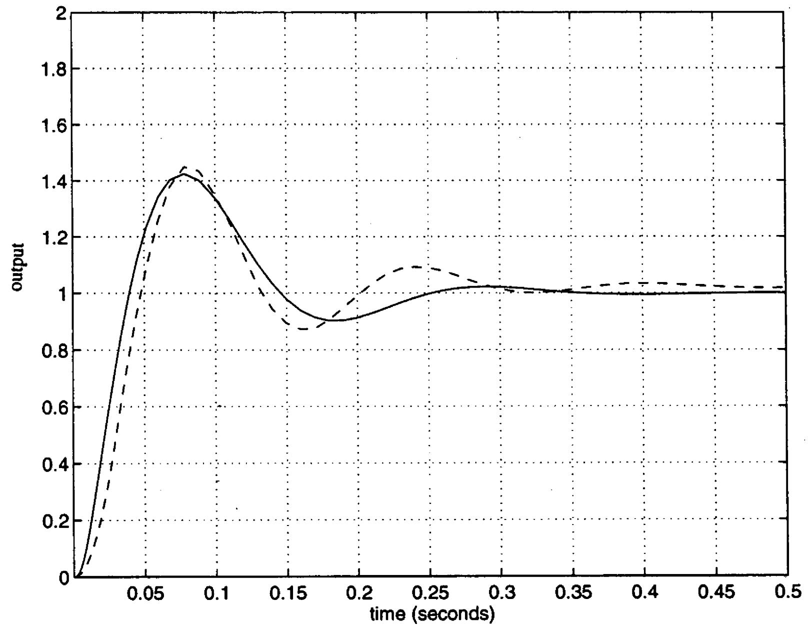

Figure 4-17: Katz example - Step response of analog system and hybrid systems with digital controller produced via the integral approximation method $T = 0.03$s

Chapter 5

Evaluation of Closed-Loop Methods

5.1 Introduction

In this chapter, methods of *closed-loop* continuous-time controller discretization that have appeared in the control literature are surveyed and compared to the methods developed in Chapter 4. Again, closed-loop discretization in this context refers to those methods which take the plant into account during the discretization. As stated in Chapter 1, methods of closed-loop controller discretization can be found in [1, 32, 43, 45, 50, 52, 53, 54, 58, 60, 61, 65, 66, 67, 71, 85]. Furthermore, the direct methods of sampled-data controller design, also mentioned in Chapter 1, can also be manipulated so that they can be used for controller discretization.

A restriction is imposed on the closed-loop discretization methods considered. The analog controller must be implementable with a unity feedback structure using a simple hold function and must have a transfer function description. This excludes discretization methods based on state-feedback controllers [52, 53, 54, 59, 85] and methods which use generalized-hold functions [68]. Multi-rate sampling is not considered.

The methods that are surveyed are those found in [1, 17, 43, 45, 58, 65] along with the three methods developed in Chapter 4. We believe that these represent a near complete cross-section of the available algorithms. The algorithm proposed by Nordström [61] is not considered. It is an L_2 based optimization scheme involving sampling, spectral factorization and linear matrix equations. It appears as if the digital controller obtained by this method has similar properties to that obtained using the H_2 optimal control approach. As pointed out in Chapter 1, the direct design of H_2 and H_∞ sampled-data controllers is possible. In this survey, the direct H_2 method found in [17] is used as a method of discretization. A method of controller discretization could also be based upon the H_∞ methods. However, it was believed that the results would mimic those of [1].

The survey proposed can be divided into three areas:

1. **Implementation:** Given that the algorithms that are surveyed have not been widely available to this point, it is important that they are easy to implement. We give a subjective analysis of each of the algorithms as far as ease of implementation is concerned.
2. **Analytical comparison:** In the work of [12] and [42], an analytical comparison of open-loop discretization methods is given. Using different methods of discretization, system properties such as bandwidth, gain and phase margin, step overshoot, e.t.c. are compared. A similar comparison is made of closed-loop discretization methods in this chapter. Along with the properties compared in [12] and [42], the order of the resulting digital controller is included.
3. **Practical comparison:** The discretization methods are also applied to a laboratory system. The system is a "two-tank apparatus" which we describe in the next section. We believe that this practical perspective is an important contribution.

This chapter is organized as follows. Section 5.2 gives a description of the two-tank apparatus used for the experimental work. An outline of the identification procedure is given. The analog control design is also outlined. In Section 5.3 the controller discretization algorithms surveyed are described. General observations are made for each algorithm. The parameter selections used with each algorithm are also given in this section. The results of the survey are presented in Section 5.4 and an analysis is undertaken in Section 5.5. Conclusions are drawn in Section 5.6.

5.2 The Apparatus

A diagrammatic representation of the two-tank apparatus is shown in Figure 5-1. For the purposes of our experiment, the voltage applied to Pump 3 (P3 - pumps water from the reservoir to Tank 1) is the control variable. Pump 1 (P1 - pumps water from Tank 1 to Tank 2) is inactive. Pump 2 (P2 - pumps water from Tank 2 to the reservoir) is generally inactive, although we use it to provide step-like disturbances as we investigate the disturbance rejection properties of the system. The valve associated with Tank 1 remains closed, whilst the valve associated with Tank 2 is set at about three quarters open. The valve between the two tanks is fully open. The height of water in both tanks can be measured via pressure transducers (TR1 and TR2), although only TR2 is used for control purposes.

Dimensions of the system are:

- Tank Size - 100 mm × 100 mm × 480 mm
- Pump Capacity[1] - P1 and P2: 500 gallons/hour, P3: 1100 gallons/hour

The control system was implemented using RTSHELL[2].

5.2.1 Identification

The method of system identification used was a least squares fit to a non-linear model. By Toricelli's Law, a model of the system represented by Figure 5-1 is given by

$$\frac{dh_1}{dt} = u - \gamma_1\sqrt{h_1 - h_2} \tag{5.2.1}$$

$$\frac{dh_2}{dt} = \gamma_1\sqrt{h_1 - h_2} - \gamma_2\sqrt{h_2} \tag{5.2.2}$$

where h_1 and h_2 are the heights in tanks one and two respectively, u is the flow rate into tank 1, and γ_1, γ_2 are constants. The rate of flow u was assumed to be a quadratic function of the voltage (V) applied to Pump 3, i.e.

$$u = \alpha_1 V^2 + \alpha_2 V + \alpha_3 \tag{5.2.3}$$

[1] Pumps produced by Rule industries inc., Gloucester MA, USA

[2] Multi Tasking Real Time Shell, developed by T. Hesketh and D.J. Clements, Department of Systems and Control, University of New South Wales, Australia

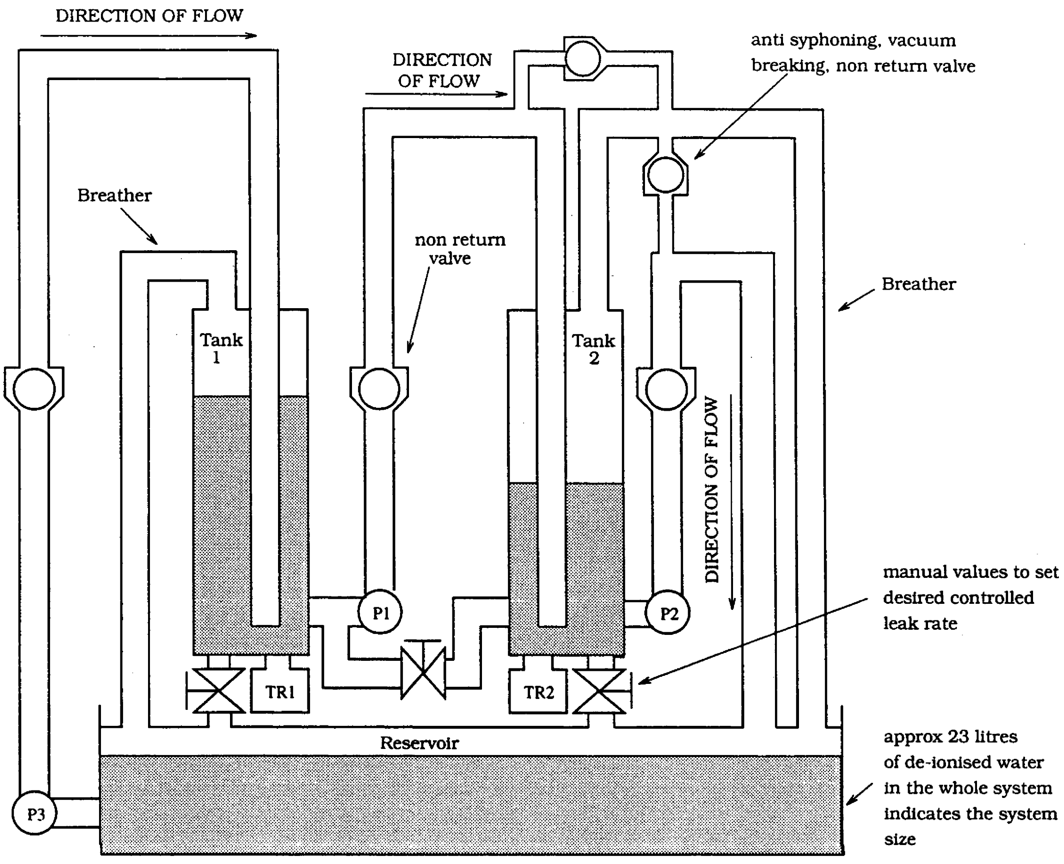
DIRECTION OF FLOW
DIRECTION OF FLOW
anti syphoning, vacuum breaking, non return valve
Breather
non return valve
Breather
Tank 1
Tank 2
DIRECTION OF FLOW
manual values to set desired controlled leak rate
P1
P2
P3
TR1
TR2
Reservoir
approx 23 litres of de-ionised water in the whole system indicates the system size

Figure 5-1: Two Tanks - Control Block Diagram

A constant term β_1 added to equation (5.2.2) yields a better fit. The system could then be described by the equations

$$\frac{dh_1}{dt} = \alpha_1 V^2 + \alpha_2 V + \alpha_3 - \gamma_1\sqrt{h_1 - h_2} \tag{5.2.4}$$

$$\frac{dh_2}{dt} = \gamma_1\sqrt{h_1 - h_2} - \gamma_2\sqrt{h_2} + \beta_1 \tag{5.2.5}$$

A series of step changes in V were used to excite the system, with data being recorded every $T = 200$ milliseconds. Approximating the derivatives by

$$\frac{dh_i(kT)}{dt} = \frac{8(h_i(kT+T) - h_i(kT-T)) - (h_i(kT+2T) - h_i(kT-2T))}{12T}, \quad i = 1,2 \tag{5.2.6}$$

yields the parameters

$$\alpha_1 = -0.9034, \ \alpha_2 = 21.0019, \ \alpha_3 = -49.3758$$

$$\beta_1 = -20.8526, \ \gamma_1 = 3.9183, \ \gamma_2 = 1.6626$$

A comparison between the simulated model and the actual system data is shown in Figure 5-2. Clearly, the result is very satisfactory.

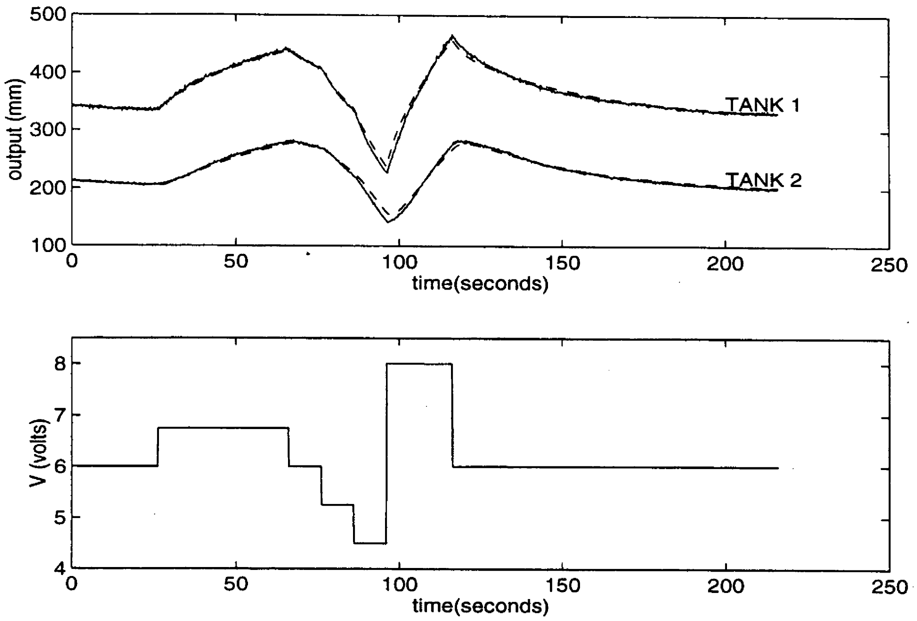

Figure 5-2: Actual system (solid line) and simulated model (dashed line)

Linearization was performed about $V_0 = 6$ V, i.e. $V = V_0 + \Delta V$. Defining

$$\Lambda_1 = \frac{\gamma_1^2}{2(\alpha_1 V_0^2 + \alpha_2 V_0 + \alpha_3)}, \; \Lambda_2 = \frac{\gamma_2^2}{\beta_1 + \alpha_1 V_0^2 + \alpha_2 V_0 + \alpha_3} \tag{5.2.7}$$

resulted in a state space model

$$\begin{bmatrix} \dot{h}_1 \\ \dot{h}_2 \end{bmatrix} = \begin{bmatrix} -\Lambda_1 & \Lambda_1 \\ \Lambda_1 & -\Lambda_1 + \Lambda_2 \end{bmatrix} \begin{bmatrix} h_1 \\ h_2 \end{bmatrix} + \begin{bmatrix} 2\alpha_1 V_0 + \alpha_2 \\ 0 \end{bmatrix} \Delta V \tag{5.2.8}$$

$$= \begin{bmatrix} -0.1740 & 0.1740 \\ 0.1740 & -0.2334 \end{bmatrix} \begin{bmatrix} h_1 \\ h_2 \end{bmatrix} + \begin{bmatrix} 10.1611 \\ 0 \end{bmatrix} \Delta V \tag{5.2.9}$$

$$h_2 = \begin{bmatrix} 0 & 1 \end{bmatrix} \begin{bmatrix} h_1 \\ h_2 \end{bmatrix} \tag{5.2.10}$$

A transfer function description between ΔV and h_2 is given by

$$\mathcal{P}(s) = \frac{1.7682}{s^2 + 0.4075s + 0.0103} \tag{5.2.11}$$

This is the nominal model which is used for the controller design.

5.2.2 Analog Controller Design

An LQG controller was designed with integral control action in order to reject constant disturbances appearing at the output. Its transfer function is given by

$$\mathcal{C}(s) = \frac{0.4014s^2 + 0.3502s + 0.0270}{s^3 + 1.7780s^2 + 2.3424s} \tag{5.2.12}$$

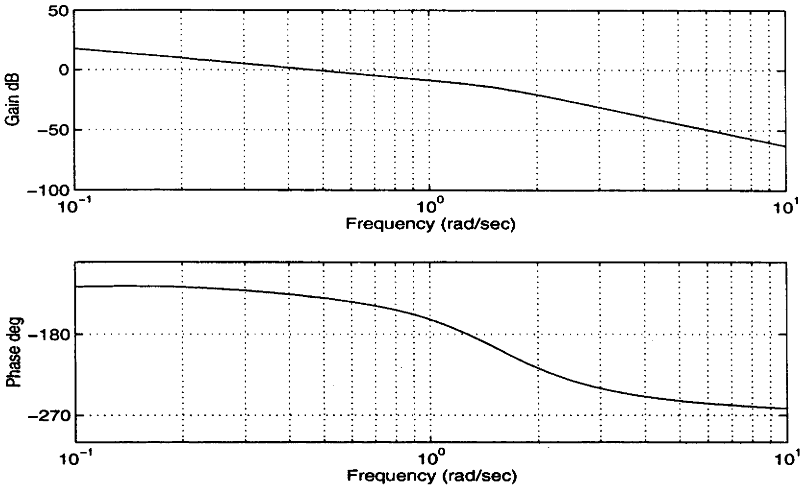

Figure 5-3: Bode Plot of Linearized System

and is implemented with a unity feedback arrangement. The Bode plot of the open-loop linearized system is displayed in Figure 5-3. The closed-loop frequency response is shown in Figure 5-4.

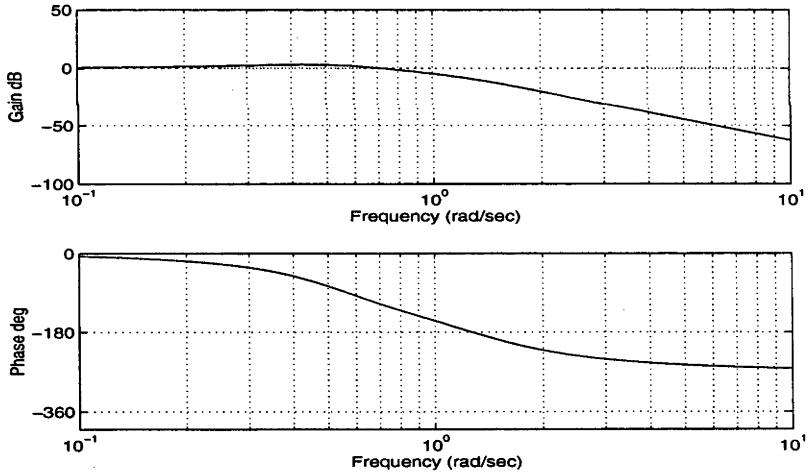

Figure 5-4: Closed-Loop Frequency Response of Linearized System

The system properties can be summarized as follows:

- Closed-loop bandwidth equal to 0.90 rad/s
- Damping coefficient of dominant poles equal to 0.51
- Gain margin equal to 11.53 dB
- Phase crossover frequency equal to 1.28 rad/s
- Phase margin equal to 41.05 degrees

- Gain Crossover frequency equal to 0.48 rad/s
- Maximum closed-loop frequency gain equal to 3.15 dB at 0.44 rad/s
- Step overshoot equal to 32%.

The sampling period for the implementation of digital controllers is chosen to be $T = 2$ seconds. This corresponds to approximately 3.5 times the closed-loop bandwidth, or approximately 6.5 times the gain crossover frequency. A rule of thumb [32] suggests that the sampling frequency be selected at least 10 times larger than the closed-loop bandwidth, so the sampling-rate used is likely to cause problems. These problems are desirable in this study, as it is well known (see [6]) that even open-loop controller discretization methods can perform reasonably well in many applications if the sampling rate is sufficiently high. Therefore, a low sampling rate that "challenges" the discretization methods is chosen.

5.3 Algorithm Descriptions

In this section, a brief description of the controller discretization methods surveyed in this chapter are given. Naturally, the interested reader should refer to the references given for a fuller description of the algorithms. The advantages and disadvantages of each algorithm are subjective. These opinions are based on our experience gained from implementing each scheme. For obvious reasons, descriptions of the algorithms developed in this thesis are omitted.

Bilinear Transformation Ben-Zwi and Preiszler in [12], perform extensive work comparing the performance of a discretized controller to the performance of an analog design. The work is restricted to open-loop methods. Katz summarizes this work in [42]. Katz makes the conclusion that the bilinear transformation is the best open-loop discretization method. It is for this reason that this method is included as a benchmark in the study of closed-loop methods.

Advantages The bilinear transformation is simple to understand and implement. The order of the analog controller is maintained.

Disadvantages In the closed-loop setting, stability not guaranteed.

Rattan's Method Rattan's algorithm [65] is a frequency-domain method of controller discretization. The method uses complex curve fitting to minimize the error between the continuous-time system and the sampled-data system. The criterion is a mean square type of the form

$$\int_{\gamma_1}^{\gamma_2} \left| F(j\gamma) - \frac{N(\gamma)}{M(\gamma)} \right|^2 \, d\gamma \tag{5.3.1}$$

where $F(\cdot)$ is the closed-loop transfer function of the analog closed-loop system. If $G_d(\cdot)$ is the closed-loop "transfer function" of the sampled-data closed-loop system, then $N(\cdot)$ and $M(\cdot)$ are the numerator and denominator of $G_d(\cdot)$ respectively. The digital controller is parameterized and substituted into (5.3.1). This equation is then differentiated with respect to the parameters and the minimum is found. The designer is responsible for the selection of γ_1 and γ_2 and also the selection of the order of the digital controller.

An extension on the basic method has also appeared in [66]. The extension takes into account the effect of the zero-order hold and any computational delays.

Advantages The designer has control over the order of the digital controller.

Disadvantages A recent paper by Hall [37] makes some comments on the methods in [65, 66]. Two main points are made. First, there are a number of errors in [66] that significantly affect its conclusions. Second, neither of the methods are guaranteed to produce a stable closed-loop system even if the original closed-loop system is stable. In fact, the two methods generally lead to unstable closed-loop system when the order of the digital compensator is large. Hall argues that the agreement between the two transfer functions is not a good measure of stability.

Our experience has confirmed these statements. Further disadvantages are that the algorithm is computationally expensive and reasonably difficult to implement. There are no guidelines for selecting the frequency range of matching (i.e. γ_1 and γ_2). Finally there are many confusing typos in [65].

Keller and Anderson's Method Keller and Anderson introduce two methods of controller discretization in [1]. The first method is *the stability margin approach.* The design of the digital controller is first framed in terms of a perturbation of the original continuous-time controller. Using a variant on the small-gain theorem [25], the authors use a measure of performance which guarantees closed-loop stability. A second method, *the closed-loop transfer function approach* basically attempts to minimize the H_∞ norm between the analog closed-loop system and the sampled-data closed-loop system.

Both methods use the techniques of "blocking" and "lifting" to approximate linear and periodically time-variant operators by discrete time-invariant operators. The digital controller is found by solving a discrete H_∞ problem. The closed-loop transfer function approach requires the designer to select an input weighting function. Both methods require the selection of N, the fast sampling rate associated with the blocking and lifting. Model order reduction is generally required.

Advantages The algorithms are easy to understand and implement. They are computationally fast and reliable. Computational delay can be accounted for.

Disadvantages The stability margin approach is basically restricted to open-loop stable controllers. The weighting function associated with the closed-loop transfer function approach can be difficult to determine—the designer requires some experience in this area. The order of the discretized controller is typically greater than the analog controller, requiring some form of controller reduction. Integral action in the controller is not necessarily preserved.

Kennedy and Evans' Method This method was one of the earliest closed-loop discretization methods developed. The motivation for this method is based on the model matching techniques presented in [6]. A series of polynomials A, B, A_m, $\tilde{B}_m$, B_e^+, B_e^-, B_p, $\tilde{B}_c$, $\tilde{B}_e$, B_e^*, $\tilde{B}_p$ are generated which are related to the hold equivalences of the plant and closed-loop system. The method necessitates the solution of a Diophantine equation. The resulting controller consists of a feedforward and a feedback part. The designer has some freedom with the selection of certain polynomials.

Advantages The order of the analog controller is maintained. The method is very suited to tracking control problems.

Disadvantages A two-degree-of-freedom controller results. The discretization algorithm is not readily implementable on a digital computer. The intersample behaviour is not considered and closed-loop properties such as disturbance rejection are not addressed. The method does not preserve controller integral action and is applicable to stable plants only.

Markazi and Hori's Method Markazi and Hori develop simple method of controller discretization in [58]. The method guarantees closed-loop stability for almost all sampling frequencies [58]. In this method, a sampled-data control system is designed such that the control input of the sampled-data control system approaches that of the continuous-time control system as the sampling frequency is increased. As with Kennedy and Evan's method, polynomials are manipulated according to certain rules and the designer is responsible for making certain selections.

Advantages The method guarantees closed-loop stability for nearly all sampling periods. It is simple to implement.

Disadvantages The resulting performance is difficult to predict. Controller integral action is not preserved.

Chen's Method In the work of Khargonekar and Sivashankar [46], Bamieh and Pearson, [9, 10] and Chen [17, 20], the sampled-data H_2 has been solved. These methods use a generalized H_2 continuous-time performance measure. The plant is given in the now popular generalized or two-port description. Strictly speaking, these methods have been developed for the direct design of sampled-data controllers. However in this survey, a generalized plant description is used which allows the controller discretization problem to be recast so that the H_2 direct method can be used. This is achieved by setting the problem up in the form of Anderson's closed-loop method. The simplified method of Chen [17] is then used. The designer must select the input weighting function.

Advantages The method is analytically attractive and numerically efficient.

Disadvantages The method is reasonably difficult to understand and to implement. The choice of norm is restrictive and the order of the resulting digital controller is large compared to that of that of the analog controller. Controller integral action is not preserved.

H_2 Optimal Control Method This method is described in Section 4.2.

Advantages The algorithm allows some flexibility in the choice of norm. Controller integral action is preserved.

Disadvantages The algorithm is reasonably complicated and difficult to implement. It is moderately computationally expensive and can suffer from numerical difficulties. The order of the discretized controller is typically greater than the analog controller.

Convex Optimization Method This method is described in Section 4.3.

Advantages This method gives the designer great flexibility in the choice of norm and allows numerous constraints to be imposed in the discretization process. The algorithm is numerically very efficient.

Disadvantages A digital controller of very high order results. A convex optimization algorithm must be available.

Integral Approximation Method This method is described in Section 4.4.

Advantages The designer has control over the order of the resulting digital controller—the order of the analog controller can be maintained. Controller integral action can be preserved.

Disadvantages The algorithm is computationally expensive due to non-linear optimization. The algorithm is fairly difficult to use because of the large number of parameters that must be selected.

All these algorithms have been coded in MATLAB. The code is summarized in Appendix F.

5.3.1 Parameter Selections

Most of the above algorithms require the selection of certain key parameters or other decisions. For a description of these, the reader is asked to refer to the associated work. In this section, the choices used for application to the two-tank experiment are summarized.

It should be noted that some of the methods produced controllers of high order. The sampled-data controller reduction procedure of [56] has been used to reduce the order so long as the performance of the controller is not significantly degraded. In all cases this resulted in digital controllers of order equal to 3, the same as the analog controller. This method reduces the order of the controller such that

$$\|W(z)[C(z) - C_r(z)]V(z)\|_\infty$$

is minimized, where $V(z)$ and $W(z)$ are weighting functions generated according to the algorithm, $C(z)$ is the original high order controller and $C_r(z)$ is the reduced order controller. The fast sampling factor N required by the controller reduction algorithm was set equal to 6.

Bilinear Transformation No selections were required.

Rattan's Method A stabilizing controller could not be found for either *3rd* or *4th* order. Many choice of γ_1 and γ_2 were tried. The *3rd* order controller found was obtained with $\gamma_1 = 0.001$ and $\gamma_2 = 200$. The numerical integration was performed via MATLAB's Adam's method.

Keller and Anderson's Method The weighting function chosen was a *3rd* order Butterworth filter with cutoff frequency $\omega_0 = \pi/(0.9T)$ cascaded with a first order transfer function with the zero at $\pi/(4T)$, the pole at -0.00001, and the gain equal to 1.4. The fast sampling factor N was set equal to 10. As a result a *6th* order controller was generated. The controller was reduced to *3rd* order.

Kennedy and Evans' Method The following polynomial selections were used:

$$\begin{aligned}
A(z) &= z^2 - 1.4145z + 0.4427\\
B(z) &= 2.7346z + 2.0861\\
A_m(z) &= z^5 - 1.2827z^4 + 0.5948z^3 - 0.1367z^2 - 0.0220z - 0.0126\\
\tilde{B}_m(z) &= 0.4062z^4 + 0.2074z^3 - 0.5045z^2 + 0.0150z + 0.0165\\
B_e^+(z) &= 1\\
B_e^-(z) &= B(z)\\
B_p(z) &= 1\\
\tilde{B}_c(z) &= (z - 0.8429)(z - 0.2108)\\
\tilde{B}_e(z) &= 0.4062(z + 1.4011)(z + 0.1632)\\
B_e^*(z) &= 0.2023(z + 0.1632)\\
\tilde{B}_p(z) &= 1
\end{aligned}$$

Markazi and Hori's Method No choices were necessary as the plant is stable.

Chen's Method The problem was set up in the same form as Anderson's method with identical weighting function. The discrete exogenous input term was introduced with $D_{21} = 0.0001$. A Kalman filter (as opposed to predictor) was used in the solution of the H_2 problem. A *21st* order controller was reduced to *3rd* order.

H_2 Optimal Control Method The reference model for a step was used with Algorithm 4.1. The resulting *8th* order controller was reduced to *3rd* order.

Convex Optimization Method The reference model for a step was used. The number of intersample points N, was set equal to 6 with $n_{sample} = 50$ and $n_{tap} = 20$. The resulting *25th* order controller was then reduced to *3rd* order. This was repeated for the l_2 norm and the l_∞ norm.

Integral Approximation Method The reference model for a step was used. The number of intersample points was equal to 4, and $\bar{\mathbf{A}} = \mathbf{0}$. All analog integrators were approximated by $1st$ order digital approximations.

5.4 Results

Each of the methods described in Section 5.3 was applied to the controller described by equation (5.2.12), using the plant information given by equation (5.2.11). The results are recorded in Table 3.1. Tables 3.2 and 3.3 summarize the system properties. A discussion of the sampled-data frequency-domain methods used is given in Appendix C. The step responses of the hybrid systems, each comprised of the nominal continuous-time plant (5.2.11) and the digital controllers derived via the various methods, are recorded in Figures 5-5 to 5-9.

Whereas Table 3.3 compares the performance of the various discretization schemes in an analytical sense, comparisons were also made when the schemes were applied to the real system. First, the digital controllers were implemented with a block cascaded with the real plant at the input side. The block enabled variations in the gain and also a delay to be introduced. The gain and then the delay were increased until the system reached a point of instability, identified by the point just before the onset of an unbounded oscillation. This procedure was not performed for a given discretization scheme if large oscillations were present without additional gain or delay. The results are summarized in Table 3.4.

Each scheme was applied to the real system and a number of tests were performed. First, the response to step changes in the reference signal was recorded. This was achieved manually through the RTSHELL program. In addition, the application of a constant output disturbance (as applied by P2) was recorded. These results are recorded in Figures 5-10 to 5-16.

DISCRETIZATION METHOD	FEEDBACK CONTROLLER	FEEDFORWARD CONTROLLER
Analog	$\frac{0.4014s^2+0.3502s+0.0270}{s^3+1.7780s^2+2.3424s}$	-
Bilinear	$\frac{0.1521z^3+0.0058z^2-0.1310z+0.0153}{z^3-0.4757z^2-0.2188z-0.3055}$	-
Rattan	$\frac{0.1769z^3+0.4961z^2+0.4610z+0.1419}{z^3+2.9730z^2+2.9443z+0.9713}$	-
Anderson and Keller	$\frac{0.1896z^3-0.0795z^2-0.1539z+0.0542}{z^3+0.0052z^2-0.7925z-0.2611}$	-
Kennedy and Evans	$\frac{0.0545z-0.0347}{z^3+0.1319z^2+0.3387z+0.1349}$	$\frac{0.2023z^3-0.1801z^2+0.0012z+0.0059}{z^3+0.1319z^2+0.3387z+0.1349}$
Markazi and Hori	$\frac{0.2313z^2-0.3114z+0.0850}{z^3-0.1506z^2-0.4305z-0.4189}$	-
Chen	$\frac{0.2643z^3-0.1357z^2-0.1276z+0.0391}{z^3+0.1390z^2-0.8106z-0.3136}$	-
H_2 optimal control	$\frac{0.1357z^3-0.0132z^2-0.1541z+0.0464}{z^3-0.1754z^2-0.5477z-0.2768}$	-
Convex (l_2)	$\frac{0.1591z^3-0.0895z^2-0.0791z+0.0194}{z^3-0.3377z^2-0.3924z-0.2698}$	-
Convex (l_∞)	$\frac{0.1546z^3+0.0175z^2-0.2089z+0.0599}{z^3+0.1947z^2-0.8811z-0.3136}$	-
Integral	$\frac{0.1629z^3-0.0651z^2-0.1185z+0.0398}{z^3-0.1892z^2-0.5956z-0.2152}$	-

Table 3.1: Controllers

PROPERTY	ABBREVIATION	UNITS
Closed-loop Bandwidth	ω_c	rad/sec
Damping Coefficient of Dominant Poles	ζ_d	-
Gain Margin	GM	dB
Phase Crossover Frequency	ω_1	rad/sec
Phase Margin	PM	degrees
Gain Crossover Frequency	ω_2	rad/sec
Maximum Closed-loop Frequency Gain	M_p	dB
Frequency at M_p	ω_3	rad/sec
Step Overshoot	σ	%

Table 3.2: Abbreviations

DISCRETIZATION METHOD	ω_c	ζ_d	GM	ω_1	PM	ω_2	M_p	ω_3	σ
Analog	0.90	0.51	11.53	1.28	41.05	0.48	3.15	0.44	32
Bilinear	0.85	0.15	3.05	0.62	15.37	0.47	12.49	0.51	91
Rattan (unstable)	0.74	0.15	-	0.59	-	0.46	10.05	0.49	58
Anderson and Keller	0.85	0.58	8.56	0.93	46.73	0.40	2.06	0.44	28
Kennedy and Evans	0.82	0.51	11.26	0.40	109.57	0.05	2.83	0.43	32
Markazi and Hori	0.89	0.51	5.06	0.60	46.89	0.28	3.24	0.45	31
Chen	0.95	0.53	9.38	1.14	41.17	0.45	3.06	0.45	31
H_2 optimal control	0.81	0.52	7.25	0.78	43.30	0.37	2.79	0.43	34
Convex (l_2)	0.80	0.48	7.87	0.87	43.19	0.38	2.73	0.42	30
Convex (l_∞)	0.80	0.52	8.14	0.84	42.75	0.38	2.80	0.41	34
Integral	0.80	0.52	8.09	0.86	42.71	0.38	2.81	0.42	34

Table 3.3: System properties

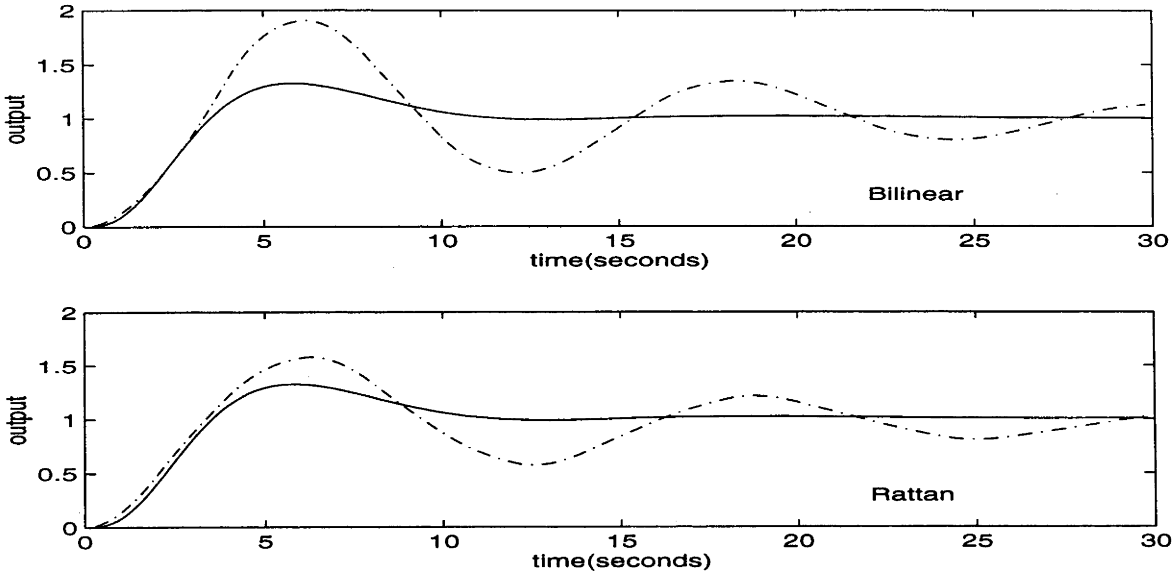

Figure 5-5: Step Response

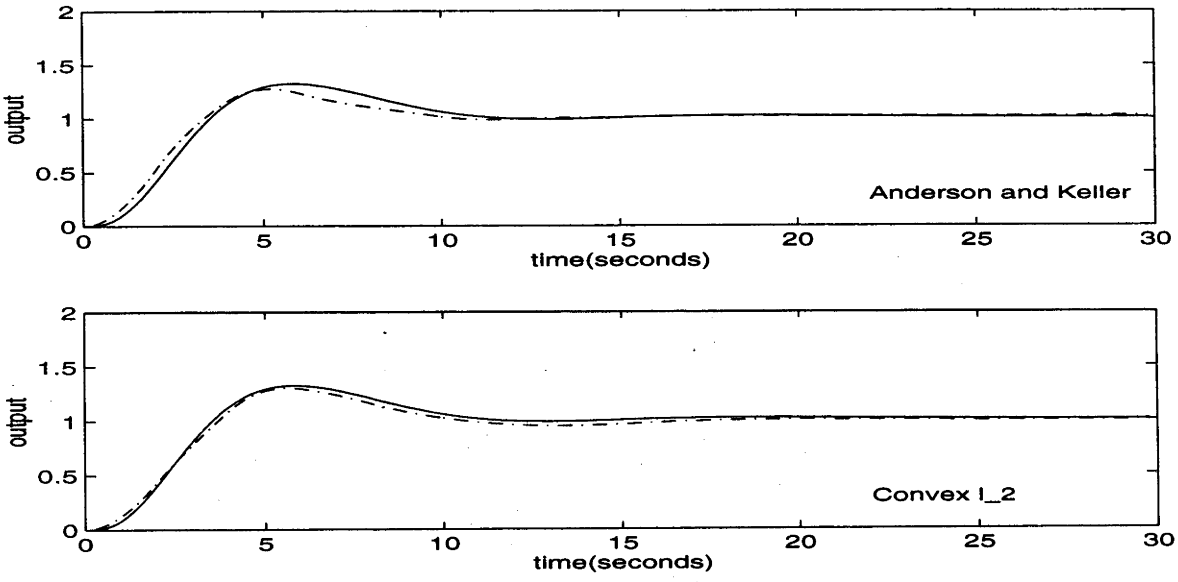

Figure 5-6: Step Response

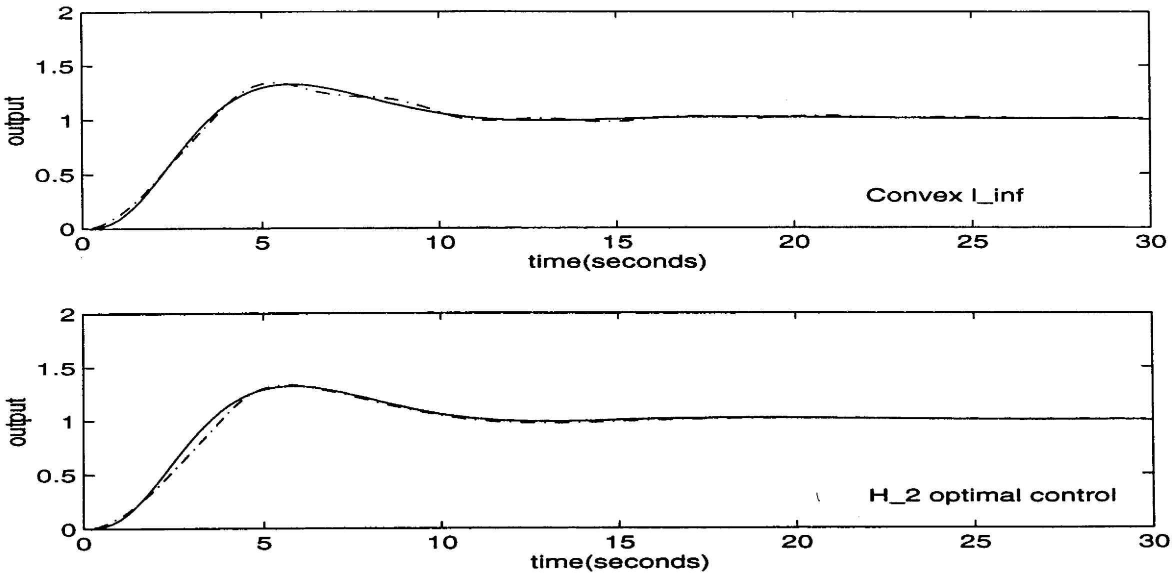

Figure 5-7: Step Response

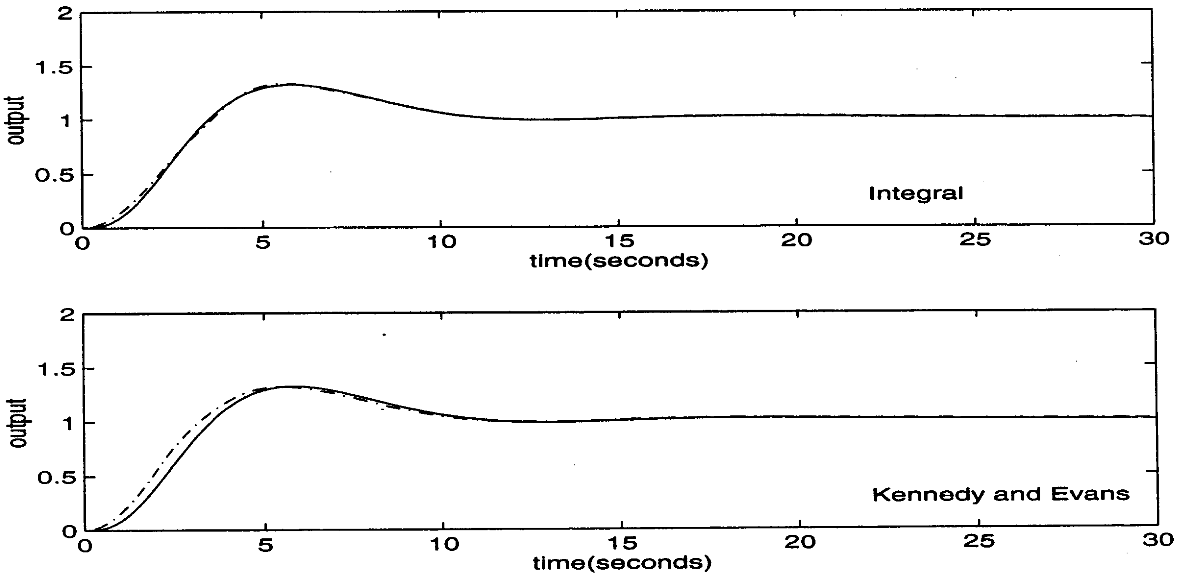

Figure 5-8: Step Response

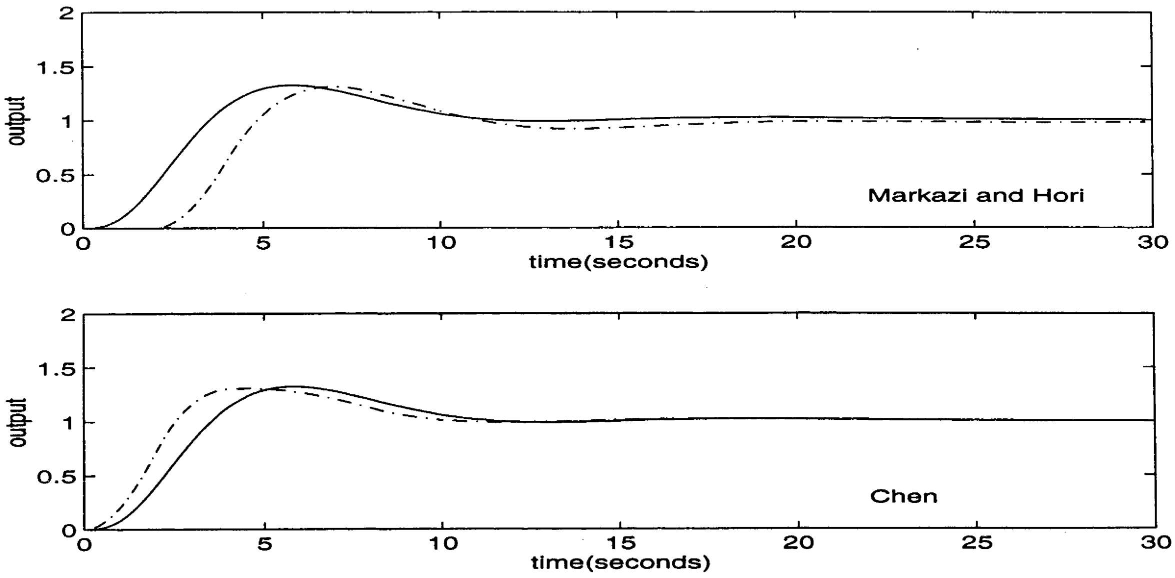

Figure 5-9: Step Response

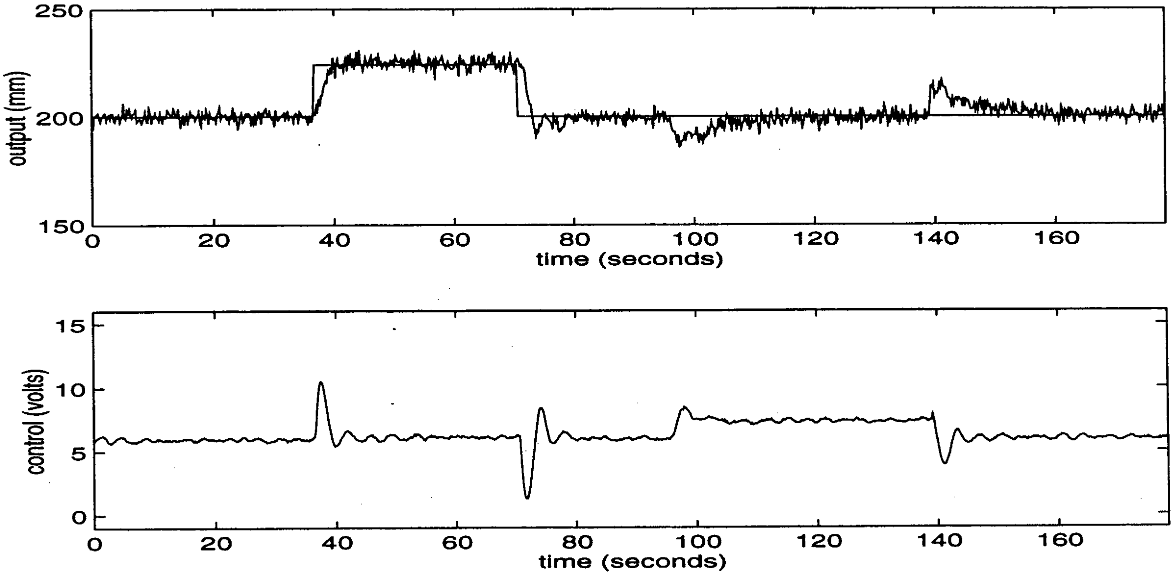

Figure 5-10: Analog Controller

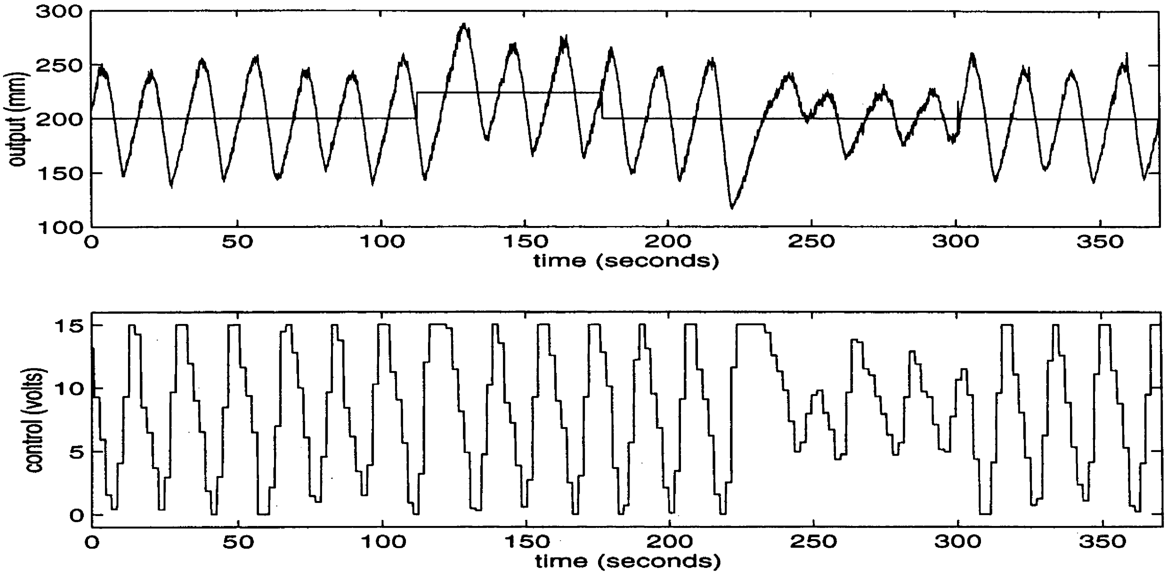

Figure 5-11: Bilinear Transformation Discrete Controller

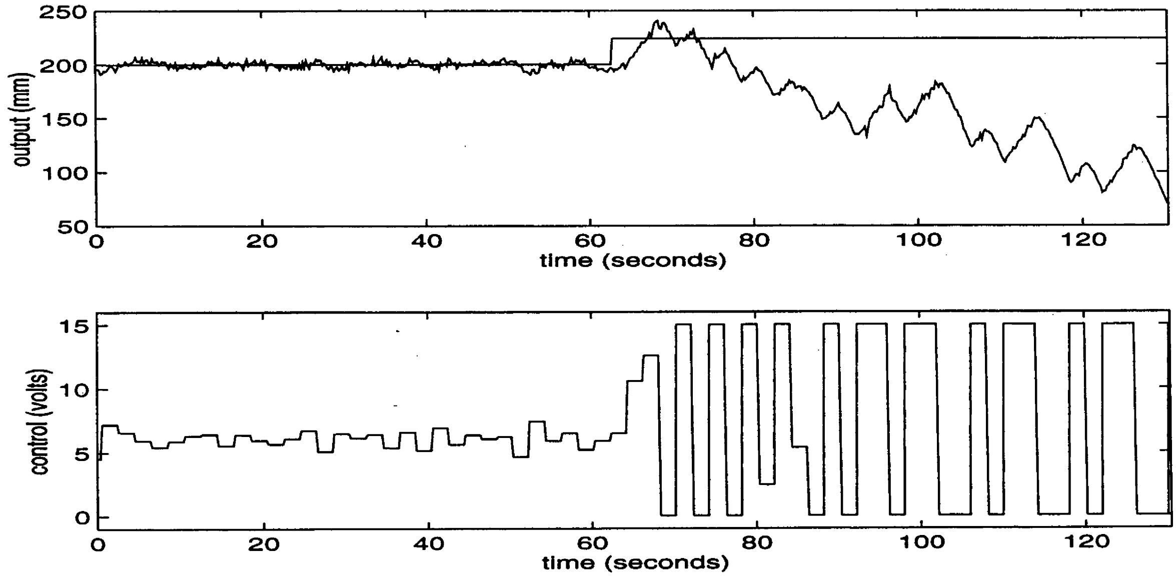

Figure 5-12: Rattan Discrete Controller

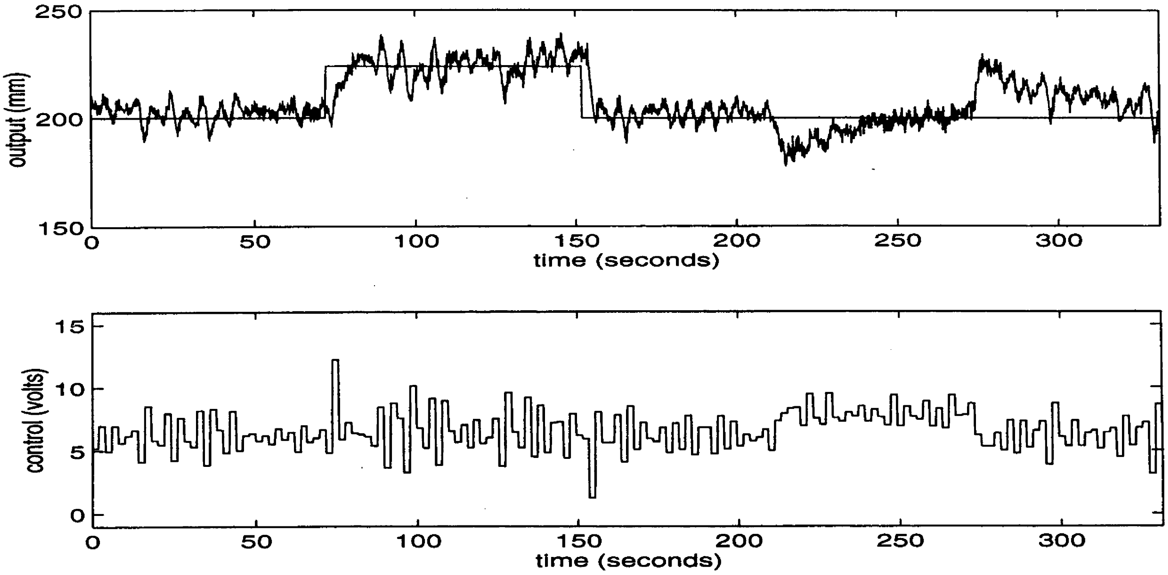

Figure 5-13: Anderson and Keller Discrete Controller

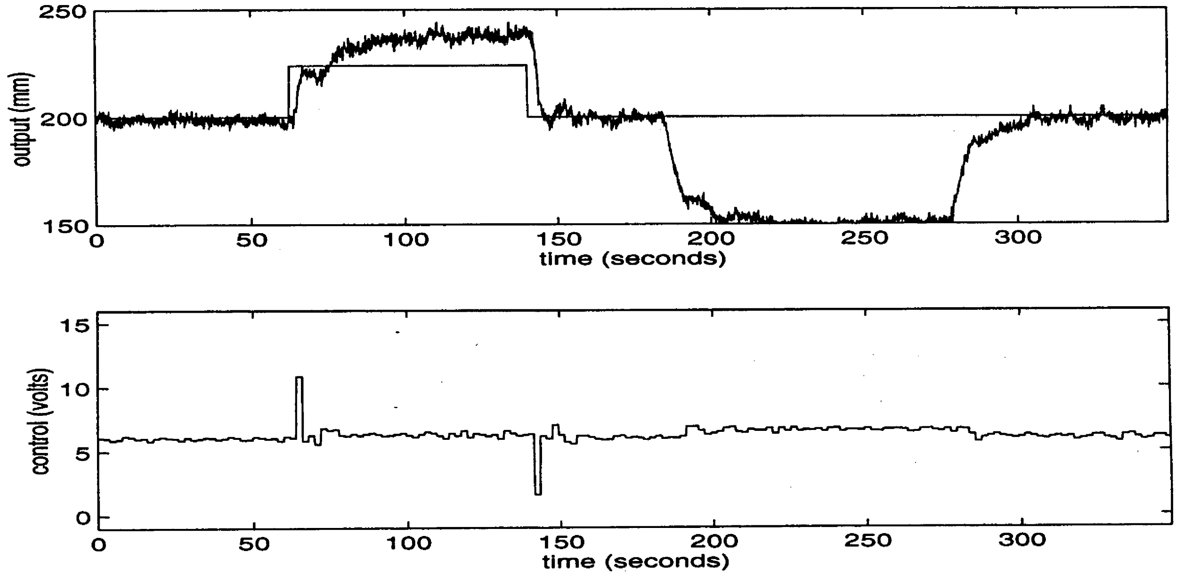

Figure 5-14: Kennedy and Evans Discrete Controller

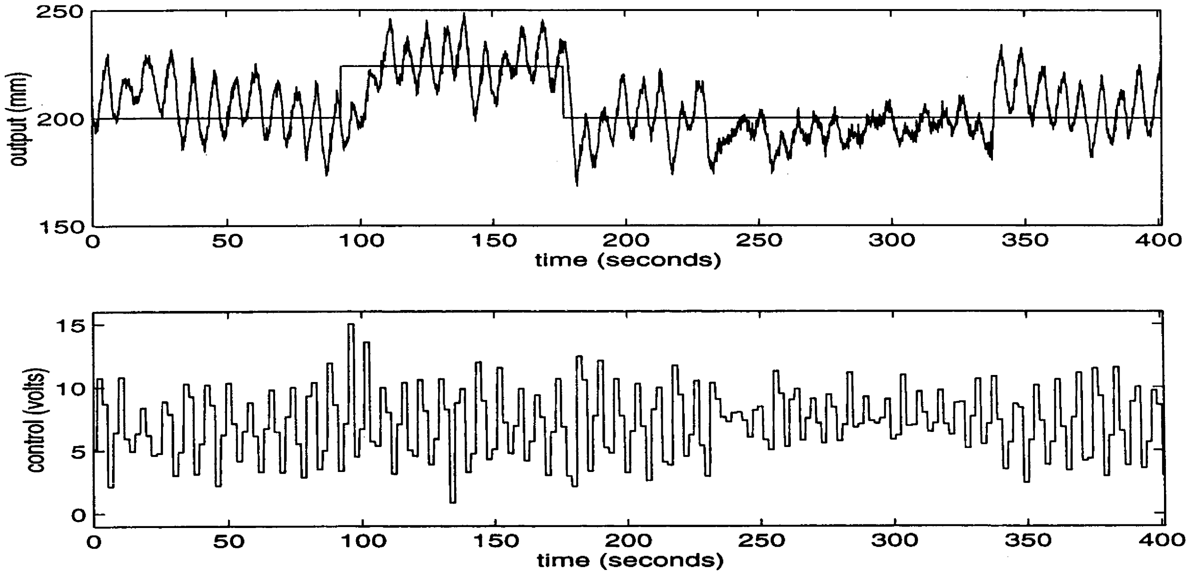

Figure 5-15: Markazi and Hori Discrete Controller

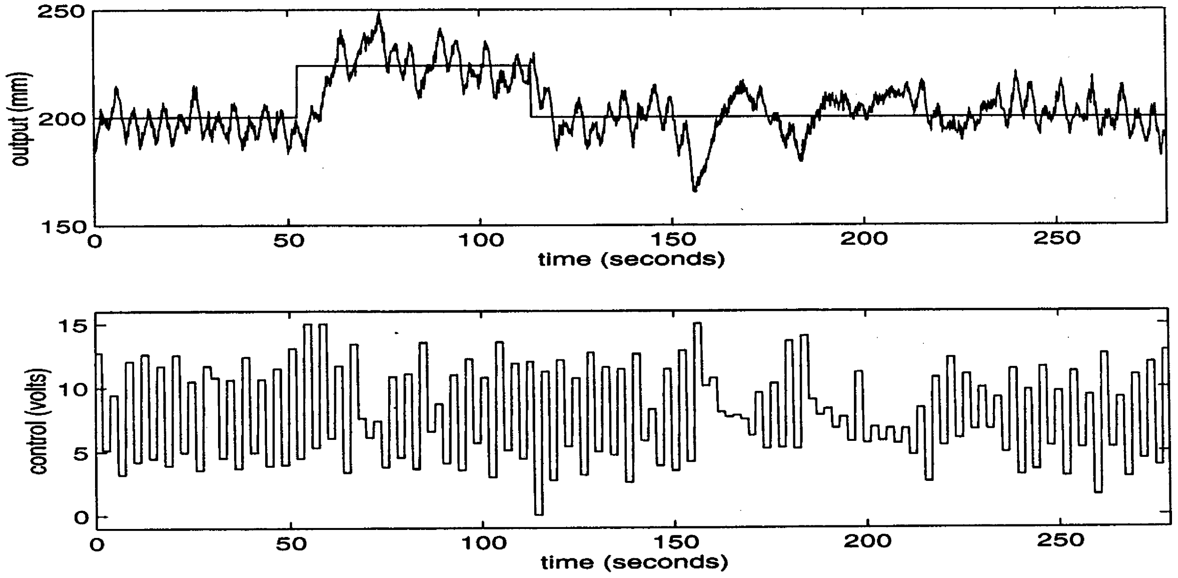

Figure 5-16: Chen Discrete Controller

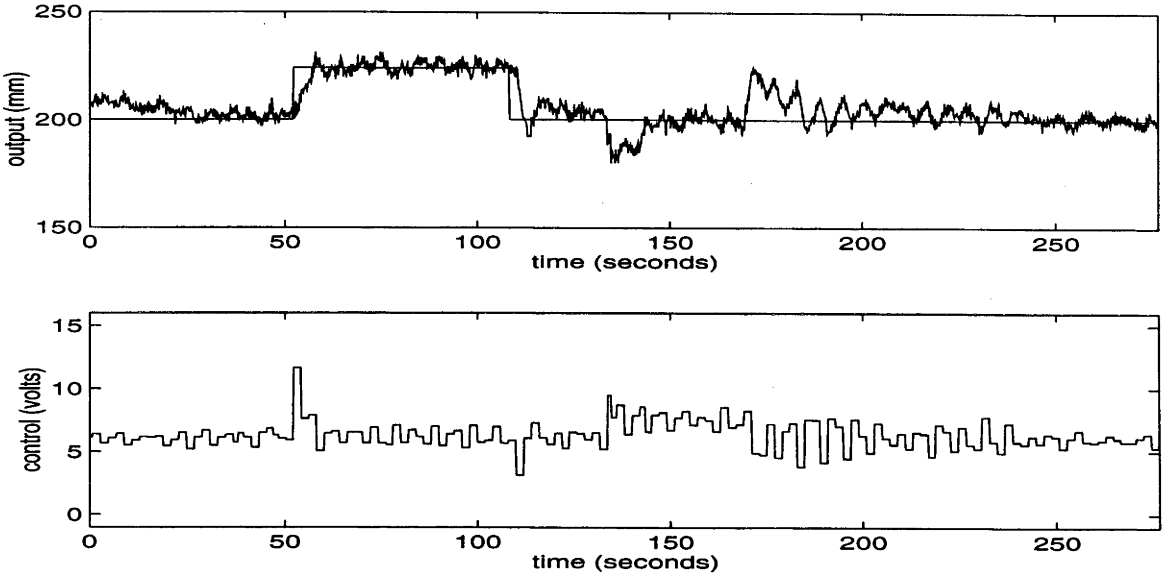

Figure 5-17: H_2 Optimal Control Discrete Controller

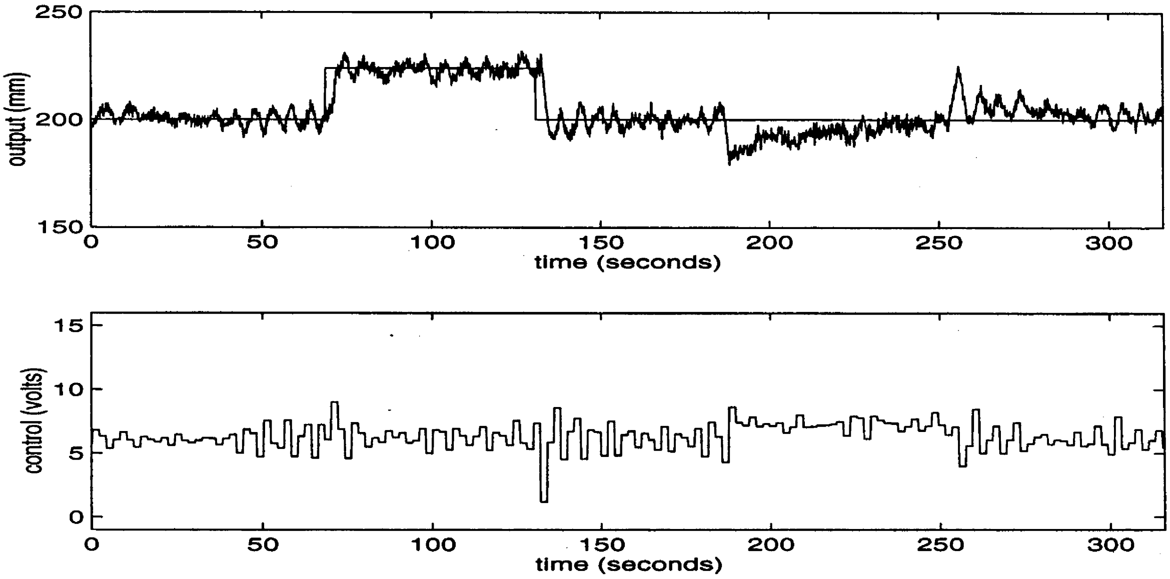

Figure 5-18: Convex Optimization l_2 Discrete Controller

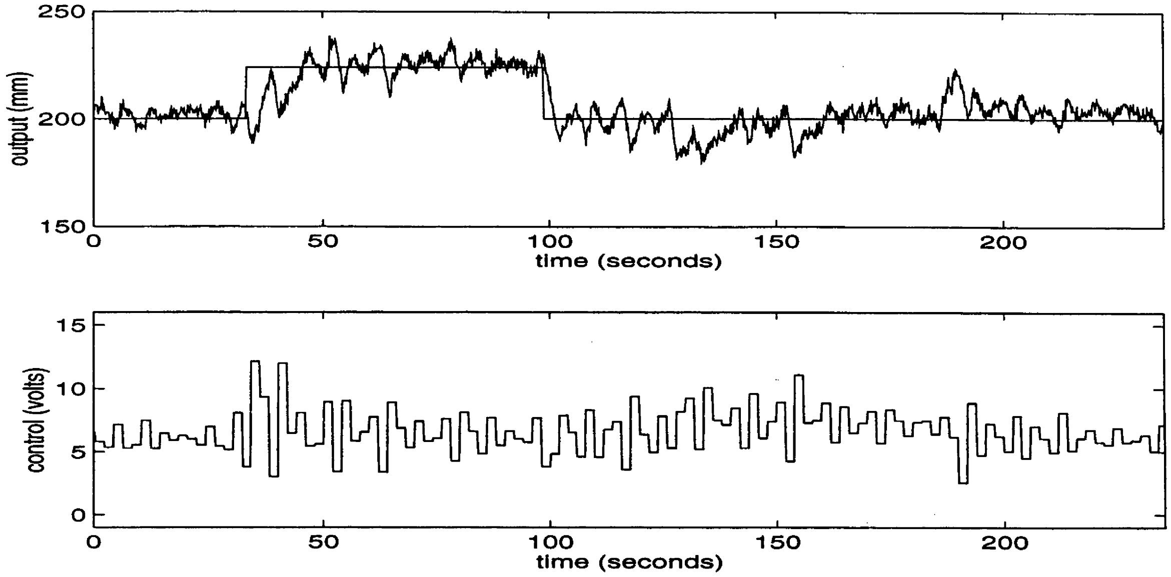

Figure 5-19: Convex Optimization l_∞ Discrete Controller

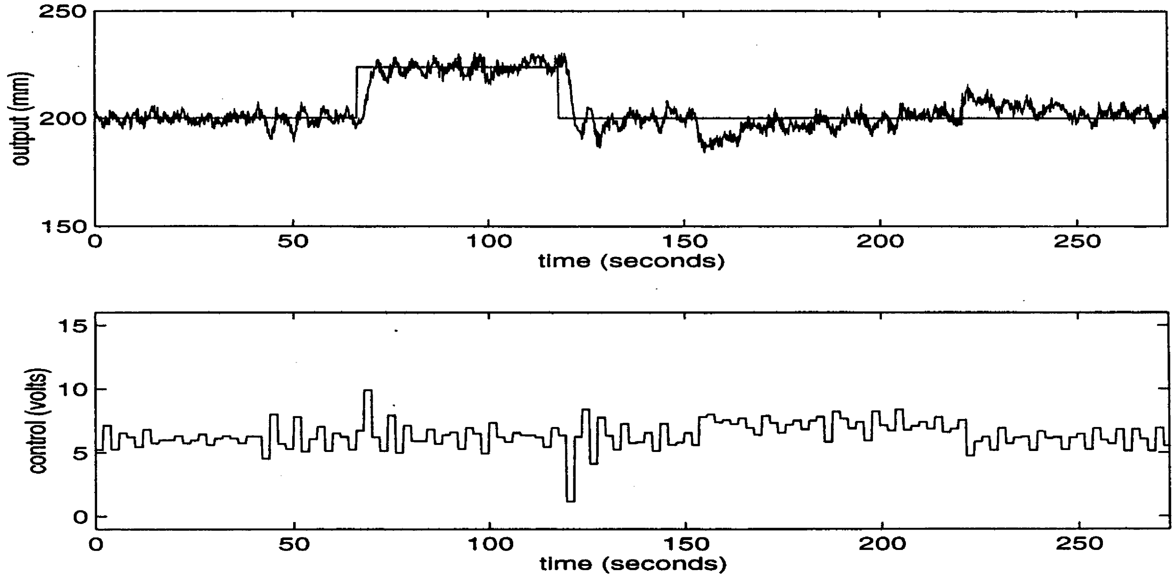

Figure 5-20: Integral Approximation Discrete Controller

DISCRETIZATION METHOD	Maximum Gain	Maximum Delay (seconds)
Analog	2.5	0.88
Bilinear	-	-
Rattan (unstable)	-	-
Anderson and Keller	1.18	1.08
Kennedy and Evans	1.03	>20
Markazi and Hori	-	-
Chen	-	-
H_2 optimal control	1.27	1.08
Convex (l_2)	1.38	0.98
Convex (l_∞)	1.26	1.32
Integral approximation	1.40	1.32

Table 3.4: System properties

5.5 Analysis

In this section, the results of Sections 5.4 are reviewed. Attempts are made to draw attention to points of interest rather than provide an extensive analysis. The reader is encouraged to examine the figures and tables of Section 5.4 more closely to verify the conclusions made.

Analog Controller The step response of the analog system model can be seen in Figures 5-5 to 5-9. In Figure 5-10, the response of the real system is shown. From time $t = 0$ to $t \approx 35$ seconds the regulating properties of the controller are shown as the system operates at an equilibrium point of 200 mm. A step is then applied such that the reference signal is changed to 224 mm. At $t \approx 70$ seconds, the reference is returned to 200 mm. At $t \approx 90$ seconds, one sees the introduction of a constant output disturbance which is removed at $t \approx 140$ seconds. The controller performs very satisfactorily throughout, with well behaved control signals. Notice that the 30% overshoot in the design is only seen in the downward reference step change.

Bilinear Transformation In Figure 5-5, the linearized model's step response is shown and the real system is shown in Figure 5-11. The performance is totally unacceptable, with wildly oscillating control signals seen in the real system (notice the scaling on the axes). This behaviour does not disappear. Notice the introduction of the constant disturbance at time $t \approx 230$ seconds.

Rattan's Method As is mentioned in Section 5.3.1, Rattan's Method is unable to find a stabilizing controller. The step response displayed in Figure 5-5 does "blowup" after sufficient time. Despite this, the controller was applied to the real system. Figure 5-12 shows the result—the unstable controller for the model is also unstable for the real system.

Keller and Anderson's Method Figures 5-6 and 5-13 show the relevant time responses of Anderson's controller. It should be noted that the "rough" appearance of the real system's response is accentuated by the scaling used. Anderson and Keller's controller has the lowest overshoot of any of the digital controllers as seen in Table 3.3. Notice the steady state offset, due to the lack of preservation of the integrator, in response to the step change in reference (Figure 5-13). The recovery after the disturbance is significantly longer than the analog design.

This method allows less additional gain in the real system than most of the other methods, as seen in Table 3.4.

Kennedy and Evans' Method Figures 5-8 and 5-14 display Kennedy and Evan's controller performance. It showed the best regulation properties around the equilibrium point. However there is a large steady state error associated with the reference step (due to lack of integral action) and the controller could not deal with the step disturbance.

Notice the extraordinary phase margin (Table 3.3) and allowable delay (Table 3.4) due to the two-degree of freedom structure. The gain margin is very large in the theoretical sense, but this is not reflected in the real system.

Markazi and Hori's Method Markazi's control scheme performance is seen in Figures 5-9 and 5-15. The delay in the step response can be seen due to the strictly proper nature of the digital controller. The controller applied to the real system has very poor performance. Notice the change from turbulent flow to laminar flow at $t \approx 250$ seconds (an unexpected disturbance!!).

The theoretical gain margin is the smallest of all controllers apart from the bilinear transformation and Rattan's method.

Chen's Method The final scheme surveyed has its results displayed in Figures 5-9 and 5-16. Chen's method has the largest bandwidth of any of the schemes (analog included). However this causes problems in the real system with very large control signals resulting.

The step response of Figure 5-9 is very revealing. Chen's method uses a norm which takes into account the fact that the input signal may not coincide with the sampling point (i.e. the input signal may come in during the intersample time). In an attempt to compensate for this, the rise time of the step response is smaller. Larger control signals are needed to achieve this. In the practical setting, the resulting digital controller output signal is wildly oscillating.

Note that the output disturbance is removed at time $t \approx 200$ seconds (possibly a little prematurely).

H_2 Optimal Control Method Figures 5-7 and 5-17 display the responses of this method. The performance is quite good, although there is a long transient time for the system to return to the equilibrium point after the step disturbance is removed.

Convex Optimization Method - l_2 The results are seen in Figures 5-6 and 5-18. Table 3.3 reveals that the damping coefficient of the dominant poles is smaller than the other methods. This is not reflected in the step overshoot of the model system, but it is seen in the real system. Disturbance attenuation is quite slow in the real system. The control signal is very well behaved in the real system.

This method is comparatively robust with respect the additional gain placed in the real system but not with respect to the time delay.

Convex Optimization Method - l_∞ Figures 5-7 and 5-19 show the time responses of the convex l_∞ optimization method. The controller applied to the real system is more "aggressive" than the l_2 controller. The step response is quite poor bearing little resemblance to the analog system. Disturbance attenuation is very good. Notice that the initial disturbance appears at $t \approx 130$ seconds. At around $t = 160$ seconds the flow through the valve at the bottom of Tank 2, changed from turbulent flow to laminar flow.

This method allows a good margin of additional delay in the real system.

Integral Approximation Method Integral approximation time responses are seen in Figures 5-8 and 5-20. This method has one of the better performances although it does take a long time to totally remove the step disturbance.

It accommodates a very good additional gain and delay in the loop.

5.6 Conclusions

This chapter has been concerned with comparing the performance of existing analog controller discretization algorithms. This comparison has been aided by practical work performance on a "two-tank" apparatus.

Perhaps a complaint that could be levelled against this work is that the dynamics of the linearized model that has been used are not sufficiently complex. This has resulted in a theoretical table of results (c.f. Table 3.3) that does not greatly differentiate between the properties of many of the resulting controllers. This objection is accepted. However, it should be added that theoretical and simulated results do not always bear witness to the performance of the controller applied to the real system. The results of this chapter support this statement.

It should be pointed out that this survey is not intended to be the final word on controller discretization. The methods applied to one practical apparatus are not sufficient to make conclusive judgements about the superiority of one scheme over another.

This said, some conclusions can be drawn. For large sampling periods, it has been demonstrated that the bilinear transformation is totally unacceptable as a method of discretization. Further, the difficulties in finding a stabilizing controller using Rattan's method suggests its ineffectiveness. The amount of numerical computation and the reasonable difficulty in implementing the scheme are also negative features. Markazi and Hori's method although valuable in finding a stabilizing controller, does not address performance issues. Kennedy and Evans' method possesses a number of positive features. However, the lack of disturbance attenuation and lack of integral action are major drawbacks. Chen's method, which is reasonably difficult to implement, produces a controller which has very large gains. In the practical setting this is a problem. The integral approximation method is one of the better methods as far as performance is concerned, but it suffers from the fact that a non-linear optimization problem must be solved. Hence, it is numerically very intensive. The H_2 optimal control and convex optimization methods also have good performance. The former method can suffer from numerical difficulties; the latter is very satisfactory from a numerical perspective. They are however, fairly difficult to understand and implement. Keller and Anderson's method has slightly inferior performance than the three methods of this author, but is easy to understand and implement. It has been found to be numerically quite a good algorithm.

Chapter 6

Conclusions

6.1 Overview of Thesis

Motivated by the requirements of modern control systems design, this thesis has considered issues associated with discretization from an engineering perspective. Two applications of discretization were identified: the open-loop problem associated with filter design, feedforward controller design, and simulation of continuous time systems; and the closed-loop problem associated with the digital re-design of a continuous-time controller. The first problem is a mature mathematical problem—dating back perhaps hundreds of years. The second problem has been motivated by technological advances and has only been considered in the last decade or so.

6.1.1 The Open-Loop Problem

While a portion of the theory presented is an extension and modification of existing mathematical techniques, the techniques developed in this thesis give important engineering insights.

The concept of signal invariant transformation has been reviewed and a comprehensive theory presented. Signal invariant transformations enable the design of digital systems whose output matches perfectly a given analog system at the sample points. If discretization is performed with the objective of minimizing the difference between the sampled output of the the analog system and the output of the discrete system, with respect to a given reference signal, then the signal invariant transformation is the optimal discretization. However the signal invariant transformation typically produces a discrete system with an order equal to the sum of the orders of the analog system and the reference model whose impulse response gives the reference signal. This thesis has addressed the problem of finding a discrete system of lower order that still produces a small discretization error with respect to the reference signal.

A parameterization of discrete-time systems was introduced by approximating each integrator of the continuous-time prototype system by a modified Newton-Cotes scheme. This parameterization was implemented in a number of different ways by selections of a matrix $\bar{\mathbf{A}}$. The order of the integral approximation could be varied so that the complexity or order of the discretization could be controlled (parallels were drawn with model order reduction techniques). An iterative optimization scheme was developed to find the parameter $\mathbf{p}$ which minimizes a "sum of squares" cost criterion.

A number of interesting results regarding the stability of the discrete-time system, the error resulting from discretization, and convexity of the cost function were presented. Interesting results were also obtained showing the dependence of the Hankel singular values (as well as the controllability and observability gramians) of the discrete-time system on the parameter $\mathbf{p}$.

A major focus of this work has been the identification of the factors which affect discretization error. By identifying these factors, the designer can determine the required complexity of the discretization. Furthermore, techniques were developed to determine the complexity of the integral approximation required for each integrator of the analog system. A scheme has been proposed which allows the designer to decide the required order of discretization as well as the order of the discrete approximation required for each integral.

The discretization method introduced in this thesis is a state space approach. Therefore the effect of the state space realization upon discretization error has been considered. A theory has been developed which is based upon the observation that parallels exist between the minimization of discretization error and the minimization of integrator noise in an analog operational amplifier circuit. A theorem has been presented which gives the state space structure required to minimize the *output noise gain* in an operational amplifier. From this, state space structures which are favourable for discretization were proposed.

Finally a large body of simulation work was undertaken which supported the above theory.

6.1.2 The Closed-Loop Problem

During the early years when digital computers were first implemented for control purposes, a simple controller discretization approach was typically used for algorithm design. A continuous-time controller was designed using the large body of available techniques and then a simple form of open-loop discretization was performed. The fact that the controller was operating in closed-loop was ignored. For small sampling periods this is not necessarily a problem, but it was discovered that the final design could be extremely unsatisfactory for larger sampling periods. Engineers realized that the closed-loop properties of the system had to be considered during the discretization phase, and as a response to this problem, a number of closed-loop discretization methods were developed.

This thesis has been concerned with developing closed-loop controller discretization methods based on optimization theory. The open-loop signal invariant transformation techniques have formed the basis of much of this theory. Three techniques have been successfully developed which were shown to perform well compared to existing methods.

The first approach discussed was based on signal invariant transformations and optimal control theory. A two-phase algorithm was developed—the first phase optimizing at a single intersample point; the second taking into account all the intersample points. Typically some form of min-max approach is taken in the second phase. This method was shown to perform well on benchmark problems. The main disadvantage with this method is the fact that a controller of large order is produced. It was shown that, if necessary, this problem could successfully be overcome by sampled-data controller reduction methods.

A second method presented was based on signal invariant transformations and convex optimization theory. This method displays a number of attractive features—flexibility in the choice of cost criterion; the ability for a variety of constraints to be incorporated into the design; and numerical reliability. Again the main disadvantage is the resulting high order of the digital controller.

The final method presented was an extension of the open-loop integral approximation method. Again signal invariant transformations were the tool used to convert the hybrid system into a pure discrete-time system which allows the solution to the problem. The motivation of this work was the fact that no model order reduction is required at the end of the design as with the other methods. The order of the digital controller is decided at the outset. This was again shown to be a very effective method of controller discretization. The main problem of this method is the resulting non-linear optimization problem that results. A technique that overcame this problem to a large extent was proposed.

An extensive comparison was made between the methods presented in this thesis and the other techniques available. This comprised a study of the methods applied to a two-tank apparatus. Comparisons were made in a theoretical and practical sense and attempts were made to give some impression of the implementational problems encountered with the algorithms.

6.2 Further Work

This section presents some open problems related to the work in this thesis.

- The H_2 optimal control method of controller discretization is very effective for achieving good performance. The major step in the algorithm involves finding the Youla parameter $\mathbf{Q}(z)$ which minimizes the l_2 model matching problem. While an elegant analytical solution is available, the numerical problem is significantly more difficult. The optimal cost is very sensitive to variations in the optimal $\mathbf{Q}(z)$. It would be very useful if effective numerical techniques were developed for this problem. Perhaps particular state space structures could be utilized. Of course this would have benefits in many applications of optimal control theory.

- Also related to the H_2 optimal control algorithm is the problem of dealing with an unstable $\mathcal{R}(s)$. As was shown the problem reduces to a two-sided interpolation problem (c.f. equation (4.2.33)). As mentioned in Chapter 4 a considerable body of knowledge exists for related problems—such as finding all the $\mathbf{W}(z) \in \mathbb{C}^{n \times m}$ such that

  $$\mathbf{x}^T(z_k)\mathbf{W}(z_k) = \mathbf{y}^T(z_k), \; k = 1, \ldots, n$$

  A starting point for the two-sided problem may be to represent the unknown matrix function in terms of a Taylor series expansion around some point. Then the collection of interpolation conditions can be represented as a system of (a-priori infinite) linear equations of the (infinitely many) Taylor coefficients. If a-priori there is no particular point where the matrix function is assumed to be analytic, then one could write the equations for Laurent coefficients.

Another point in relation to the unstable reference model case is in relation to the procedure proposed in Chapter 4. The final optimization involves a convex optimization over matrix functions and constant matrices. An efficient numerical scheme would be useful for this problem. This would have benefits in other fields, for instance there are problems in robotics which require this form of optimization.

- Again in relation to the H_2 optimal control method is the issue relating to the relative degree of the reference model $\mathcal{R}(s)$. A sufficient condition for the realizability of the digital controller is that the relative degree of all the components of $\mathcal{R}(s)$ equal one (c.f. Section 4.2.2). This implies, for example, that the method is suitable for a cosine reference model, but not a sine reference model. This is intuitively puzzling. It would be interesting to understand this problem more fully.

- A more general problem of discretization methods for non-linear systems needs to be addressed. At this point closed-loop discretization schemes for non-linear systems do not exist. In many applications, for instance in robotics applications, the dynamics of the system to be controlled may be highly non-linear. Such a system may be controlled using either a linear controller or a non-linear controller. Given that the digital computer is becoming more widely used for control tasks, the need for effective discretization is apparent.

  It would be interesting to consider if some form of non-linear signal invariant transformation theory exists.

Appendix A

Factorization Theory

Suppose $[\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D}]$ is a minimal realization of a transfer function matrix

$$\mathbf{G}(z) = \mathbf{D} + \mathbf{C}(z\mathbf{I}_{n_x} - \mathbf{A})^{-1}\mathbf{B} \in \mathbb{C}^{n_1 \times n_2}$$

which has no poles at $z = 0$. A realization of $\mathbf{G}^{-}(z)$ is given by

$$[-(\mathbf{A}^{-1})^T, (\mathbf{C}\mathbf{A}^{-1})^T, -(\mathbf{A}^{-1}\mathbf{B})^T, (\mathbf{D} - \mathbf{C}\mathbf{A}^{-1}\mathbf{B})^T] \qquad \text{(A.0.1)}$$

$\mathbf{G}(z)$ is said to be *inner* if

$$\mathbf{G}^{-}(z)\mathbf{G}(z) = \mathbf{I}_{n_2}$$

and *co-inner* if

$$\mathbf{G}(z)\mathbf{G}^{-}(z) = \mathbf{I}_{n_1}$$

A stable function $\mathbf{G}(z)$ (i.e. $\mathbf{G}(z) \in RH^{\infty}$) is said to be *outer* if $\mathbf{G}(z)$ has full row rank n_1, $\forall |z| > 1$. A stable function $\mathbf{G}(z)$ is said to be *co-outer* if $\mathbf{G}(z)$ has full column rank n_2, $\forall |z| > 1$.

Given any $\mathbf{G}(z) \in RH^{\infty}$, the factorization

$$\mathbf{G}(z) = \mathbf{G}_i(z)\mathbf{G}_o(z)$$

where $\mathbf{G}_i(z)$ is inner and $\mathbf{G}_o(z)$ is outer is said to be an *inner-outer* factorization. The factorization

$$\mathbf{G}(z) = \mathbf{G}_{co}(z)\mathbf{G}_{ci}(z)$$

where $\mathbf{G}_{co}(z)$ is co-outer and $\mathbf{G}_{ci}(z)$ is co-inner is said to be an *co-inner-co-outer* factorization.

Theorem A.1 *Consider the $n_1 \times n_2$ $(n_1 \geq n_2)$ transfer function matrix*

$$\mathbf{G}(z) = \mathbf{D} + \mathbf{C}(z\mathbf{I}_{n_x} - \mathbf{A})^{-1}\mathbf{B} \in RH^{\infty}$$

Suppose

(a). $\mathbf{G}(z)$ *has no poles at* $z = 0$.

(b). $\mathbf{G}^{-}(z)\mathbf{G}(z) > \mathbf{0}$; $z = e^{j\theta}$; $\forall \theta \in [0, 2\pi]$

(c). $\mathbf{G}^{-}(z)\mathbf{G}(z)|_{z=\infty}$ *has full rank* n_2

(d). $\mathbf{D}^T\mathbf{D} > \mathbf{0}$

Then $\mathbf{G}(z)$ *has an inner-outer factorization*

$$\mathbf{G}(z) = \mathbf{G}_i(z)\mathbf{G}_o(z)$$

where

$$\begin{aligned} \mathbf{G}_i(z) &\triangleq \mathbf{D}\mathbf{M} + (\mathbf{C} + \mathbf{D}\mathbf{L})(z\mathbf{I}_{n_x} - (\mathbf{A} + \mathbf{B}\mathbf{L}))^{-1}\mathbf{B}\mathbf{M} \\ \mathbf{G}_o(z) &\triangleq \mathbf{M}^{-1} - \mathbf{M}^{-1}\mathbf{L}(z\mathbf{I}_{n_x} - \mathbf{A})^{-1}\mathbf{B} \end{aligned}$$

and

$$\begin{aligned} \mathbf{M} &\triangleq (\mathbf{D}^T\mathbf{D} + \mathbf{B}^T\mathbf{X}\mathbf{B})^{-1/2} \\ \mathbf{L} &\triangleq -(\mathbf{D}^T\mathbf{D} + \mathbf{B}^T\mathbf{X}\mathbf{B})^{-1}(\mathbf{D}^T\mathbf{C} + \mathbf{B}^T\mathbf{X}\mathbf{A}) \end{aligned}$$

and where $\mathbf{X}$ *is the stabilizing solution of the discrete algebraic Riccati equation*

$$\mathbf{X} = \tilde{\mathbf{A}}^T\mathbf{X}\tilde{\mathbf{A}} - \tilde{\mathbf{A}}^T\mathbf{X}\tilde{\mathbf{B}}(\mathbf{I}_{n_2} + \tilde{\mathbf{B}}^T\mathbf{X}\tilde{\mathbf{B}})^{-1}\tilde{\mathbf{B}}^T\mathbf{X}\tilde{\mathbf{A}} + \mathbf{Q}$$
$$\tilde{\mathbf{A}} \triangleq \mathbf{A} - \mathbf{B}(\mathbf{D}^T\mathbf{D})^{-1}\mathbf{D}^T\mathbf{C} \; ; \; \tilde{\mathbf{B}} \triangleq \mathbf{B}(\mathbf{D}^T\mathbf{D})^{-1/2}$$
$$\mathbf{Q} \triangleq \mathbf{C}^T(\mathbf{I}_{n_1} - \mathbf{D}(\mathbf{D}^T\mathbf{D})^{-1}\mathbf{D}^T)\mathbf{C}$$

Proof: See [36] ▮

An analogous result for finding a co-inner-co-outer factorization exists.

Appendix B

Stability Theory for Sampled-Data Systems

In this appendix, a summary of stability theory for sampled-data systems is given. The material presented is based on the treatment in [19, 29].

First define a number of *linear* operators. The *sampling operator* $\mathbf{S}_T : \mathcal{C}^0([0,\infty),\mathbb{C}^n) \to \mathcal{S}(\mathbb{Z}^+,\mathbb{C}^n)$ is defined by

$$\mathbf{y} = \mathbf{S}_T\mathbf{u} \Leftrightarrow \mathbf{y}(0) = \mathbf{0},\ \mathbf{y}(k) = \mathbf{u}(kT) \text{ for } k \geq 1 \tag{B.0.1}$$

Similarly define the *zero-order hold operator* $\mathbf{H}_T : \mathcal{S}(\mathbb{Z}^+,\mathbb{C}^n) \to \mathcal{C}^0([0,\infty),\mathbb{C}^n)$ by

$$\mathbf{y} = \mathbf{H}_T\mathbf{u} \ \Leftrightarrow \mathbf{y}(0) = \mathbf{u}(0),\ \mathbf{y}(t) = \mathbf{u}(k) \text{ for } kT < t \leq (k+1)T \tag{B.0.2}$$

The *backward shift operator* $\mathbf{U} : \mathcal{S}(\mathbb{Z}^+,\mathbb{C}^n) \to \mathcal{S}(\mathbb{Z}^+,\mathbb{C}^n)$ is defined by

$$\mathbf{y} = \mathbf{U}\mathbf{u} \Leftrightarrow \mathbf{y}(0) = \mathbf{0},\ \mathbf{y}(k) = \mathbf{u}(k-1) \text{ for } k \geq 1 \tag{B.0.3}$$

Similarly the adjoint operator, the *forward shift operator* $\mathbf{U}^* : \mathcal{S}(\mathbb{Z}^+,\mathbb{C}^n) \to \mathcal{S}(\mathbb{Z}^+,\mathbb{C}^n)$ is defined by

$$\mathbf{y} = \mathbf{U}\mathbf{u} \Leftrightarrow \mathbf{y}(0) = \mathbf{0},\ \mathbf{y}(k) = \mathbf{u}(k+1) \text{ for } k \geq 0 \tag{B.0.4}$$

Consider the sampled-data system depicted in Figure B-1 where the linear time-invariant system $\mathbf{P}_c$ has a state model

$$\dot{\mathbf{x}}_c = \mathbf{A}_c\mathbf{x}_c + \mathbf{B}_c\mathbf{u} \tag{B.0.5}$$

$$\mathbf{y} = \mathbf{C}_c\mathbf{x}_c + \mathbf{D}_c\mathbf{u} \tag{B.0.6}$$

and the digital controller K has a state model

$$\boldsymbol{\xi}(k+1) = \mathbf{A}_k\boldsymbol{\xi}(k) + \mathbf{B}_k\boldsymbol{\nu}(k) \tag{B.0.7}$$

$$\boldsymbol{\psi}(k) = \mathbf{C}_k\boldsymbol{\xi}(k) + \mathbf{D}_k\boldsymbol{\nu}(k) \tag{B.0.8}$$

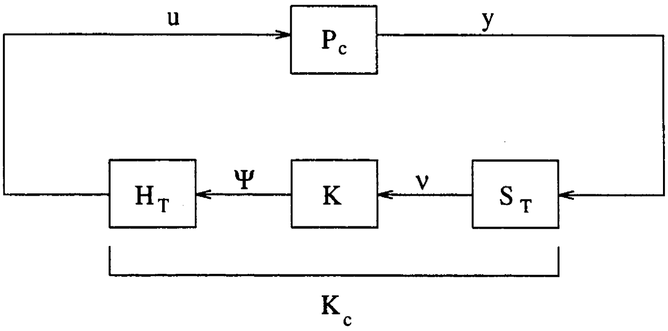

Figure B-1: Sampled-data system

The authors of [29] propose that a state for the sampled-data system can be given by

$$\mathbf{x}_{hybrid} = \begin{bmatrix} \mathbf{x}_c \\ \mathbf{H}_T\boldsymbol{\psi} \\ \mathbf{H}_T\mathbf{U}^*\boldsymbol{\xi} \end{bmatrix} \tag{B.0.9}$$

A state at time t gives sufficient information to compute all future values of all signals, given the present and future inputs. It is straightforward to show that (B.0.9) gives this information, and hence that $\mathbf{x}_{hybrid}$ is a state for the sampled-data system.

By moving the sampling operator and the zero-order hold operator around the loop in Figure B-1, a discrete-time feedback system is formed as in Figure B-2.

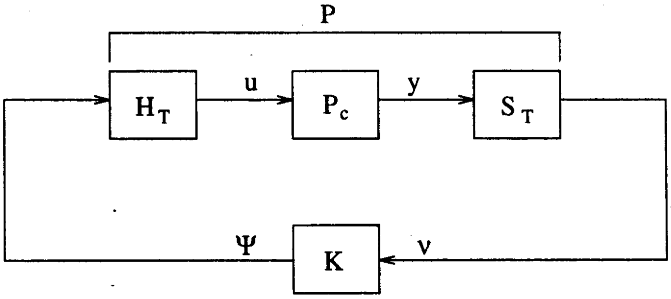

Figure B-2: Discrete-time system

Define $\mathbf{P} \triangleq \mathbf{S}_T\mathbf{P}_c\mathbf{H}_T$. According to (B.0.5), $\mathbf{P}$ has a discrete-time state model given by

$$\mathbf{x}_c(kT+T) = e^{\mathbf{A}_cT}\mathbf{x}_c(kT) + \int_0^T e^{\mathbf{A}_c\tau}\,d\tau\,\mathbf{B}_c\boldsymbol{\psi}(k) \tag{B.0.10}$$

$$\boldsymbol{\nu}(k) = \mathbf{C}_c\mathbf{x}_c(kT) + \mathbf{D}_c\boldsymbol{\psi}(k-1) \tag{B.0.11}$$

With $\mathbf{D}_c = \mathbf{0}$ this is a standard state model, however in general ($\mathbf{D}_c \neq \mathbf{0}$) the state of $\mathbf{P}$ at time k is given by

$$\begin{bmatrix} \mathbf{x}_c(kT) \\ \psi(k-1) \end{bmatrix} \tag{B.0.12}$$

with state parameters

$$\mathbf{A}_s \triangleq \begin{bmatrix} e^{\mathbf{A}_c T} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \quad ; \quad \mathbf{B}_s \triangleq \begin{bmatrix} \int_0^T e^{\mathbf{A}_c \tau}\, d\tau\, \mathbf{B}_c \\ \mathbf{I}_n \end{bmatrix}$$
$$\mathbf{C}_s \triangleq \begin{bmatrix} \mathbf{C}_c & \mathbf{D}_c \end{bmatrix} \quad ; \quad \mathbf{D}_s \triangleq \mathbf{0}$$

Of interest is the relationship between the stability of discrete-time feedback system of Figure B-2 and that of the hybrid feedback system of Figure B-1.

Lemma B.1 *Assume $\mathbf{A}_c$ (c.f. B.0.5) satisfies the following*

1. *whenever μ is an eigenvalue of $\mathbf{A}_c$ with nonnegative real part, none of the points* $\mu + j2\pi k/h$, $k \neq 0$, is an eigenvalue of $\mathbf{A}_c$
2. *none of the points $j2\pi k/h$, $k \neq 0$, is an eigenvalue of $\mathbf{A}_c$*

Then, if $(\mathbf{A}_c, \mathbf{B}_c)$ is stabilizable in continuous-time, so is $(\mathbf{A}_s, \mathbf{B}_s)$ in discrete-time. Similarly, if $(\mathbf{C}_c, \mathbf{A}_c)$ is detectable in continuous-time, so is $(\mathbf{C}_s, \mathbf{A}_s)$ in discrete-time.

The conditions of this lemma ensure that the discretization of the loop does not introduce unstabilizable or undetectable modes.

Definition B.1 (Discrete-time) *The state* $\mathbf{x}$ converges exponentially *if, for zero input, there exists positive constants α and β such that, for every initial time k_0 and initial state* $\mathbf{x}(k_0)$

$$\|\mathbf{x}(k)\| \leq \|\mathbf{x}(k_0)\| \beta e^{-\alpha(k-k_0)}, \; k \geq k_0 \tag{B.0.13}$$

Definition B.2 (Continuous-time) *The state* $\mathbf{x}_{hybrid}$ converges exponentially *if, for zero input, there exists positive constants α and β such that, for every initial time t_0*

and initial state $\mathbf{x}_{hybrid}(t_0)$

$$\|\mathbf{x}_{hybrid}(t)\| \leq \|\mathbf{x}_{hybrid}(t_0)\|\beta e^{-\alpha(t-t_0)},\ t \geq t_0 \tag{B.0.14}$$

If $\mathbf{x}_{P;K}$ is the sampled hybrid state, i.e. $\mathbf{x}_{P;K}(k) = (\mathbf{S}_T\mathbf{x}_{hybrid})(k)$ then the following result relates the stability of a hybrid system and that of the corresponding discrete-time system.

Theorem B.1 *Assume* $\mathbf{A}_c$ *satisfies the conditions of Lemma B.1. If* $\mathbf{x}_{P;K}$ *is exponentially convergent, so is* $\mathbf{x}_{hybrid}$.

Under the same conditions as Lemma B.1, it can be shown that the hybrid system is L_∞ input-output stable. However the authors of [19] note that L_2 input-output stability is not guaranteed, due to the fact that the sampler does not map every continuous function in L_2 into l_2. This problem can be alleviated by introducing a strictly causal stable continuous-time filter prior to the sampler.

Appendix C

Sampled-Data Frequency Response Methods

Consider Figure C-1 where a continuous-time signal r is sampled, processed by a digital controller $C(z)$, and then applied to a continuous-time process $P(s)$ via a zero order hold.

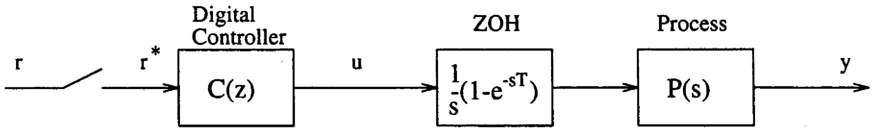

Figure C-1: Block diagram of sampler, digital controller, zero order hold, and continuous-time process

Let $R(s), R^*(s), U(s), Y(s)$ be the Laplace transforms of the signals r, r^*, u, y respectively. The notation $R^*(s)$ is used to symbolize the Laplace transform of $r^*(t)$, the sampled or impulse-modulated $r(t)$. It is well known ([32]) that

$$R^*(s) = \frac{1}{T} \sum_{n=-\infty}^{\infty} R(s - jn\omega_s) \tag{C.0.1}$$

where T is the sampling period and ω_s is the sampling frequency in radians per second ($\omega_s = 2\pi T^{-1}$). Also

$$C(z) = C^*(s)|_{e^{sT}=z} \tag{C.0.2}$$

The continuous frequency response of the sampled-data system is given by

$$\frac{1}{T} \sum_{n=-\infty}^{\infty} \frac{1}{s}(1 - e^{-sT})P(s)\ C(z)|_{z=e^{sT}}\Big|_{s=j\omega_n} \tag{C.0.3}$$

where $\omega_n = \omega + 2\pi n T^{-1}$, $(n = \dots, -2, -1, 0, 1, 2, \dots)$. The $n = 0$ term corresponds to the fundamental component, and is equivalent to

$$C(z)\tilde{P}(z)|_{z=e^{j\omega T}} \tag{C.0.4}$$

where $\tilde{P}(z)$ is the discrete hold equivalent of $P(s)$. For $n \neq 0$, the frequency response for the individual alias components are obtained. In this thesis, the quantity (C.0.4) is used as an approximation to (C.0.3). The approximation is a very good one, with significant deviations occurring near the Nyquist frequency only.

As a final note, Gessing [33] suggests the utilization of Bode plots of the control index [28] for frequency analysis. This plot is the magnitude of the frequency transfer function from the reference signal to the error signal in a classic unity feedback configuration. This is not used in Chapter 5 as it adds little to the analysis.

Appendix D

Derivation of Error Bounds

D.1 Proof of Lemma 3.9

The integral solution of equation (3.6.2) is given by

$$\begin{aligned} x(kT+T) &= e^{-aT}x(kT) + \int_{kT}^{kT+T} e^{-a(kT+T-\sigma)}bu(\sigma)\,d\sigma \qquad \text{(D.1.1)} \\ &= e^{-aT}x(kT) + \int_{0}^{T} e^{-at}bu(kT+T-t)\,dt \qquad \text{(D.1.2)} \end{aligned}$$

In the case where the integral is approximated by the trapezoidal rule ($\alpha = \beta = \frac{T}{2}$), one has

$$\begin{aligned} x(kT+T) &= e^{-aT}x(kT) + \frac{bT}{2}[u(kT+T) + e^{-aT}u(kT)] \\ &- \frac{T^3b}{12}\frac{d^2}{dt^2}\{e^{-at}u(kT+T-t)\}|_{t=\xi_k} \quad 0 < \xi_k < T \qquad \text{(D.1.3)} \end{aligned}$$

The last term in equation (D.1.3) is the error term associated with the trapezoidal rule. If $\hat{x}(kT)$ is the approximation of $x(kT)$ at time kT, then the error at time kT is defined by $E(kT)$, where

$$E(kT) \triangleq x(kT) - \hat{x}(kT)$$

By iteration, one can show that

$$E(kT) = \frac{-bT^3}{12}\sum_{j=0}^{k-1} e^{-aT(k-j-1)}\frac{d^2}{dt^2}\{e^{-at}u((j+1)T-t)\}|_{t=\xi_k} \quad 0 < \xi_j < T \qquad \text{(D.1.4)}$$

Calculating the derivatives gives

$$\begin{aligned} E(kT) = \frac{-bT^3}{12}\sum_{j=0}^{k-1} e^{-aT(k-j-1)}\{a^2 e^{-a\xi_j} u((j+1)T-t) \\ +2ae^{-a\xi_j} u((j+1)T-t) + e^{-a\xi_j} u((j+1)T-t)\} \end{aligned} \qquad \text{(D.1.5)}$$

It follows that

$$|E(kT)| \leq \frac{|b|T^3}{12}\sum_{j=0}^{k-1} e^{-aT(k-j-1)}\{a^2|u_{max}| + |2au'_{max}| + |u''_{max}|\} \qquad \text{(D.1.6)}$$

Define

$$E = \sup_k |E(kT)|$$

then

$$\begin{aligned} E &\leq \frac{|b|T^3}{12}\left|\frac{1}{1-e^{-aT}}\right|\{a^2|u_{max}| + |2au'_{max}| + |u''_{max}|\} && \text{(D.1.7)} \\ &= \frac{1}{12}\left|\frac{b}{a}\right| T^2\{a^2|u_{max}| + |2au'_{max}| + |u''_{max}|\} + O(T^3) && \text{(D.1.8)} \end{aligned}$$

Given the input $u(t) = U\cos(\omega t)$,

$$\begin{aligned} E &\leq \frac{1}{12}\left|\frac{b}{a}\right| T^2\{a^2 U + 2aU\omega + U\omega^2\} + O(T^3) && \text{(D.1.9)} \\ &= \frac{1}{12}\left|\frac{b}{a}\right| U(\omega T + aT)^2 + O(T^3) && \text{(D.1.10)} \end{aligned}$$

The error at the output can be written as

$$\begin{aligned} E_{output} &\leq \frac{1}{12}\left|\frac{cb}{a}\right| U(\omega T + aT)^2 + O(T^3) && \text{(D.1.11)} \\ &= \frac{1}{6}\sigma U(\omega T + aT)^2 + O(T^3) && \text{(D.1.12)} \end{aligned}$$

where σ is the Hankel singular value.

D.2 Proof of Lemma 3.10

Using the fact that

$$e^{\mathbf{A}t} = e^{-a_1 t}\begin{bmatrix} \cos(a_2 t) & \sin(a_2 t) \\ -\sin(a_2 t) & \cos(a_2 t) \end{bmatrix}$$

and defining

$$E_{output}(kT) = y(kT) - \hat{y}(kT)$$

it can be shown that

$$\begin{aligned}
E_{output}(kT) &= \frac{-T^3}{12}\begin{bmatrix} 1 & 0 \end{bmatrix}\sum_{j=0}^{k-1} e^{-a_1T(k-j-1)} \\
&\times \begin{bmatrix} \cos(a_2T(k-j-1)) & \sin(a_2T(k-j-1)) \\ -\sin(a_2T(k-j-1)) & \cos(a_2T(k-j-1)) \end{bmatrix} \\
&\times \begin{bmatrix} \frac{d^2}{dt^2}\{e^{-a_1t}\sqrt{b_1^2+b_2^2}\sin(a_2t+\arctan(\frac{b_1}{b_2}))u(kT+T-t)\}|_{t=\xi_j^1} \\ \frac{d^2}{dt^2}\{e^{-a_1t}\sqrt{b_1^2+b_2^2}\sin(a_2t-\arctan(\frac{b_2}{b_1}))u(kT+T-t)\}|_{t=\xi_j^2} \end{bmatrix}
\end{aligned}$$

from which it can be shown that for $u(t) = U\cos(\omega t)$,

$$\begin{aligned}
|E_{output}(kT)| &\leq \frac{T^3}{12}U\sqrt{b_1^2+b_2^2}\,(\omega+(a_1+|a_2|))^2\sum_{j=0}^{k-1} e^{-\alpha T(k-j-1)} \\
&\times \{|\cos(a_2T(k-j-1))|+|\sin(a_2T(k-j-1))|\}
\end{aligned}$$

This bound is not heavily dependent upon β. Taking the supremum over k gives

$$E_{output} \leq \frac{1}{3}U\frac{\sqrt{b_1^2+b_2^2}}{2\alpha}(\omega T+(a_1+|a_2|)T)^2 \qquad \text{(D.2.1)}$$

D.3 Proof of Lemma 3.11

A fast sampled system of sample period $\frac{T}{N}$, $N \to \infty$ is generated according to Lemma 3.3. Using "blocking" techniques the fast sampled system becomes

$$\begin{aligned}
\mathbf{x}_F(kT+T) &= \mathbf{F}_p^N\mathbf{x}_F(kT) + \begin{bmatrix} \mathbf{F}_p^{N-1}\mathbf{G}_p & \mathbf{F}_p^{N-2}\mathbf{G}_p & \ldots & \mathbf{G}_p \end{bmatrix}\Xi\tilde{\mathbf{u}} && \text{(D.3.1)} \\
y_F(kT) &= \mathbf{H}_p\mathbf{x}_F(kT) + \begin{bmatrix} J_p & 0 & \ldots & 0 \end{bmatrix}\Xi\tilde{\mathbf{u}} && \text{(D.3.2)}
\end{aligned}$$

where $\tilde{\mathbf{u}}$ is given by

$$\tilde{\mathbf{u}} = \begin{bmatrix} u(kT) \\ u(kT+\frac{T}{N}) \\ \cdot \\ \cdot \\ \cdot \\ u(kT+(N-1)\frac{T}{N}) \end{bmatrix} \qquad \text{(D.3.3)}$$

and Ξ is a matrix which determines the correlation of the elements of $\tilde{\mathbf{u}}$. One would expect that, if a sensible choice of sampling period has been made, then the elements of the vector $\tilde{\mathbf{u}}$ would be in some sense correlated for a given N. The correlation matrix of $\tilde{\mathbf{u}}$ is given by

$$\mathbf{R}_u = \Xi\Xi^T = \begin{bmatrix} 1 & \lambda & \lambda^2 & \cdot & \cdot & \cdots & \lambda^{N-1} \\ \lambda & 1 & \lambda & \lambda^2 & \cdot & \cdots & \lambda^{N-2} \\ \lambda^2 & \lambda & 1 & \lambda & \lambda^2 & \dots & \lambda^{N-3} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdots & \cdot \\ \lambda^{N-1} & \lambda^{N-2} & \cdot & \cdot & \cdot & \cdots & 1 \end{bmatrix}, \qquad \text{(D.3.4)}$$

where $\lambda = e^{-\frac{\varsigma T}{N-1}}$. The constant ς is a quantity which determines the amount of correlation between the members of $\tilde{\mathbf{u}}$, with smaller correlation as $\varsigma \to \infty$. Performing a Cholesky decomposition on $\mathbf{R}_u$ shows that one choice of Ξ is

$$\Xi = \begin{bmatrix} 1 & 0 & 0 & \cdot & \cdot & \cdots & 0 \\ \lambda & \sqrt{1-\lambda^2} & 0 & 0 & \cdot & \cdots & 0 \\ \lambda^2 & \lambda\sqrt{1-\lambda^2} & \sqrt{1-\lambda^2} & 0 & 0 & \cdots & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdots & \cdot \\ \lambda^{N-1} & \lambda^{N-2}\sqrt{1-\lambda^2} & \cdot & \cdot & \cdot & \cdots & \sqrt{1-\lambda^2} \end{bmatrix} \qquad \text{(D.3.5)}$$

The discrete system with period T is given by

$$\mathbf{x}_S(kT+T) = \mathbf{F}_p^N \mathbf{x}_S(kT) + \begin{bmatrix} \mathbf{F}_p^{N-1}\mathbf{G}_p + \mathbf{F}_p^{N-2}\mathbf{G}_p + \cdots + \mathbf{G}_p & \mathbf{0} & \cdots & \mathbf{0} \end{bmatrix} \Xi\tilde{\mathbf{u}} \qquad \text{(D.3.6)}$$

$$y_S(kT) = \mathbf{H}_p \mathbf{x}_\mathbf{S}(kT) + \begin{bmatrix} J_p & 0 & \cdots & 0 \end{bmatrix} \Xi\tilde{\mathbf{u}} \qquad \text{(D.3.7)}$$

A state space description of $G_e(z, \mathbf{p}, N)$ is then given by

$$\begin{aligned} \begin{bmatrix} \mathbf{x}_S(kT+T) \\ \mathbf{x}_F(kT+T) \end{bmatrix} &= \Phi \begin{bmatrix} \mathbf{x}_\mathbf{S}(kT) \\ \mathbf{x}_\mathbf{F}(kT) \end{bmatrix} + \Gamma\Xi\tilde{\mathbf{u}} \\ e(kT) = y_S(kT) - y_F(kT) &= \begin{bmatrix} \mathbf{H}_p & -\mathbf{H}_p \end{bmatrix} \begin{bmatrix} \mathbf{x}_\mathbf{S}(kT) \\ \mathbf{x}_\mathbf{F}(kT) \end{bmatrix} \end{aligned} \qquad \text{(D.3.8)}$$

where

$$\Phi \triangleq \begin{bmatrix} \mathbf{F}_p^N & \mathbf{0} \\ \mathbf{0} & \mathbf{F}_p^N \end{bmatrix} = \begin{bmatrix} e^{\mathbf{A}T} & \mathbf{0} \\ \mathbf{0} & e^{\mathbf{A}T} \end{bmatrix} \text{ and}$$

$$\Gamma \triangleq \begin{bmatrix} \mathbf{F}_p^{N-1}\mathbf{G}_p & \mathbf{F}_p^{N-2}\mathbf{G}_p & \dots & \mathbf{G}_p \\ \mathbf{F}_p^{N-1}\mathbf{G}_p + \mathbf{F}_p^{N-2}\mathbf{G}_p + \cdots + \mathbf{G}_p & \mathbf{0} & \dots & \mathbf{0} \end{bmatrix}$$

The H_2 norm of the system (D.3.8) as $N \to \infty$ is now computed. This requires the solution of the Lyapunov equation

$$\mathbf{P}_N = \mathbf{\Phi}\mathbf{P}_N\mathbf{\Phi}^T + \mathbf{\Gamma}\mathbf{R}_u\mathbf{\Gamma}^T \tag{D.3.9}$$

in the limit as $N \to \infty$. When the product $\mathbf{\Gamma\Xi}$ is calculated, it can be seen that each element in the matrix takes the form of a geometric series. Summing each series gives

$$\mathbf{\Gamma\Xi} = \left[\begin{matrix} (\lambda\mathbf{I}_n - \mathbf{F}_p)^{-1}(\lambda^N\mathbf{I}_n - \mathbf{F}_p^N)\mathbf{G}_p & \sqrt{1-\lambda^2}(\lambda\mathbf{I}_n - \mathbf{F}_p)^{-1}(\lambda^{N-1}\mathbf{I}_n - \mathbf{F}_p^{N-1})\mathbf{G}_p \\ (\mathbf{I}_n - \mathbf{F}_p)^{-1}(\mathbf{I}_n - \mathbf{F}_p^N)\mathbf{G}_p & \mathbf{0} \end{matrix}\right.$$
$$\left.\begin{matrix} \sqrt{1-\lambda^2}(\lambda\mathbf{I}_n - \mathbf{F}_p)^{-1}(\lambda^{N-2}\mathbf{I}_n - \mathbf{F}_p^{N-2})\mathbf{G}_p & \dots & \dots & \sqrt{1-\lambda^2}\mathbf{G}_p \\ \mathbf{0} & \dots & \dots & \mathbf{0} \end{matrix}\right] \tag{D.3.10}$$

The Lyapunov equation is analytic in its arguments so the limit may be taken before the formation of the infinite sum. Thus, in the case $\boldsymbol{\alpha} + \boldsymbol{\beta} = T\mathbf{I}_n$, it can be shown that

$$\mathbf{\Gamma} \to \begin{bmatrix} e^{-\varsigma T}\int_0^T e^{(\mathbf{A}+\varsigma\mathbf{I}_n)\tau}\,d\tau\,\mathbf{b} & \mathbf{0} & \dots & \mathbf{0} \\ \int_0^T e^{A\tau}\,d\tau\,\mathbf{b} & \mathbf{0} & \dots & \mathbf{0} \end{bmatrix} \text{ as } N \to \infty$$

The solution to the Lyapunov equation (D.3.9) in infinite sum form is

$$\mathbf{P}_\infty = \sum_{k=0}^{\infty} \left[\begin{matrix} e^{k\mathbf{A}T}e^{-2\varsigma T}(\int_0^T e^{(\mathbf{A}+\varsigma\mathbf{I}_n)\tau}\,d\tau)\mathbf{b}\mathbf{b}^T(\int_0^T e^{(\mathbf{A}+\varsigma\mathbf{I}_n)\tau}\,d\tau)^T e^{k\mathbf{A}^T T} \\ e^{k\mathbf{A}T}e^{-\varsigma T}(\int_0^T e^{\mathbf{A}\tau}\,d\tau)\mathbf{b}\mathbf{b}^T(\int_0^T e^{(\mathbf{A}+\varsigma\mathbf{I}_n)\tau}\,d\tau)^T e^{k\mathbf{A}^T T} \end{matrix}\right. \tag{D.3.11}$$
$$\left.\begin{matrix} e^{k\mathbf{A}T}e^{-\varsigma T}(\int_0^T e^{(\mathbf{A}+\varsigma\mathbf{I}_n)\tau}\,d\tau)\mathbf{b}\mathbf{b}^T(\int_0^T e^{\mathbf{A}\tau}\,d\tau)^T e^{k\mathbf{A}^T T} \\ e^{k\mathbf{A}T}(\int_0^T e^{\mathbf{A}\tau}\,d\tau)\mathbf{b}\mathbf{b}^T(\int_0^T e^{\mathbf{A}\tau}\,d\tau)^T e^{k\mathbf{A}^T T} \end{matrix}\right] \tag{D.3.12}$$

Thus

$$
\begin{aligned}
&||G_e(z, \mathbf{p}, \infty)||_2^2 \\
= \;& \sum_{k=0}^{\infty} \mathbf{c}^T e^{k\mathbf{A}T} \{ e^{-2\varsigma T} (\int_0^T e^{(\mathbf{A}+\varsigma\mathbf{I}_n)\tau}\, d\tau) \mathbf{b}\mathbf{b}^T (\int_0^T e^{(\mathbf{A}+\varsigma\mathbf{I}_n)\tau}\, d\tau)^T \\
+ \;& (\int_0^T e^{\mathbf{A}\tau}\, d\tau) \mathbf{b}\mathbf{b}^T (\int_0^T e^{\mathbf{A}\tau}\, d\tau)^T - e^{-\varsigma T} (\int_0^T e^{\mathbf{A}\tau}\, d\tau) \mathbf{b}\mathbf{b}^T (\int_0^T e^{(\mathbf{A}+\varsigma\mathbf{I}_n)\tau}\, d\tau)^T \\
- \;& e^{-\varsigma T} (\int_0^T e^{(\mathbf{A}+\varsigma\mathbf{I}_n)\tau}\, d\tau) \mathbf{b}\mathbf{b}^T (\int_0^T e^{\mathbf{A}\tau}\, d\tau)^T \} e^{k\mathbf{A}^T T} \mathbf{c} \\
= \;& \sum_{k=0}^{\infty} \mathbf{c}^T e^{k\mathbf{A}T} [\int_0^T e^{\mathbf{A}\tau}\, d\tau - e^{-\varsigma T} (\int_0^T e^{(\mathbf{A}+\varsigma\mathbf{I}_n)\tau}\, d\tau)] \mathbf{b}\mathbf{b}^T \\
\times \;& [\int_0^T e^{\mathbf{A}\tau}\, d\tau - e^{-\varsigma T} (\int_0^T e^{(\mathbf{A}+\varsigma\mathbf{I}_n)\tau}\, d\tau)]^T e^{k\mathbf{A}^T T} \mathbf{c}
\end{aligned}
\tag{D.3.13}
$$

$$
\leq \; T \sum_{k=0}^{\infty} \mathbf{c}^T e^{k\mathbf{A}T} \int_0^T [e^{\mathbf{A}\tau} - e^{\mathbf{A}\tau + \varsigma\mathbf{I}_n(\tau - T)}] \mathbf{b}\mathbf{b}^T [e^{\mathbf{A}\tau} - e^{\mathbf{A}\tau + \varsigma\mathbf{I}_n(\tau - T)}]^T \, d\tau \; e^{k\mathbf{A}^T T} \mathbf{c}
\tag{D.3.14}
$$

$$
\begin{aligned}
= \;& T \sum_{k=0}^{\infty} \mathbf{c}^T e^{k\mathbf{A}T} \int_0^T \{1 - e^{\varsigma(\tau - T)}\}^2 e^{\mathbf{A}\tau} \mathbf{b}\mathbf{b}^T e^{\mathbf{A}^T\tau} \, d\tau \; e^{k\mathbf{A}^T T} \mathbf{c} \\
\leq \;& T\{1 - e^{-\varsigma T}\}^2 \sum_{k=0}^{\infty} \mathbf{c}^T \int_0^T e^{k\mathbf{A}T + \mathbf{A}\tau} \mathbf{b}\mathbf{b}^T e^{k\mathbf{A}^T T + \mathbf{A}^T\tau} \, d\tau \; \mathbf{c} \\
= \;& T\{1 - e^{-\varsigma T}\}^2 \mathbf{c}^T \int_0^\infty e^{\mathbf{A}\tau} \mathbf{b}\mathbf{b}^T e^{\mathbf{b}\mathbf{A}^T\tau} \, d\tau \; \mathbf{c} \\
= \;& T\{1 - e^{-\varsigma T}\}^2 ||\mathcal{G}(s)||_2^2
\end{aligned}
\tag{D.3.15}
$$

Inequality (D.3.14) follows by application of the Cauchy-Schwartz inequality.

Appendix E

The Continuous-Discrete Lyapunov Equation

Consider the equation

$$\gamma \mathbf{A}\mathbf{X}\mathbf{A}^T + \mathbf{A}\mathbf{X} + \mathbf{X}\mathbf{A}^T + \kappa \mathbf{Q} = \mathbf{0} \quad \mathbf{Q} = \mathbf{Q}^T \geq 0 \tag{E.0.1}$$

for $\gamma \neq 0$, which shall be called the Continuous-Discrete Lyapunov Equation because of its similarity to both types of Lyapunov equations. Assume $\mathbf{A} \in \mathbb{R}^{n \times n}$, and denote the linear space of all $n \times n$ real valued matrices, with the usual rules of multiplication and addition, by $(\mathcal{X}, \mathbb{R})$. The dimension of $(\mathcal{X}, \mathbb{R})$ is n^2. Corresponding to (E.0.1) introduce the operator $\boldsymbol{\mathcal{Z}}$

$$\boldsymbol{\mathcal{Z}} : \ (\mathcal{X}, \mathbb{R}) \rightarrow (\mathcal{X}, \mathbb{R}) \text{ where } \boldsymbol{\mathcal{Z}}(\mathbf{X}) \triangleq \gamma \mathbf{A}\mathbf{X}\mathbf{A}^T + \mathbf{A}\mathbf{X} + \mathbf{X}\mathbf{A}^T \quad \forall \mathbf{X} \in \mathcal{X} \tag{E.0.2}$$

Consider

$$\mathbf{X} = \boldsymbol{\xi}_i \boldsymbol{\xi}_j^T \tag{E.0.3}$$

where $\boldsymbol{\xi}_i$ is an eigenvector of $\mathbf{A}$; i.e.

$$\mathbf{A}\boldsymbol{\xi}_i = \lambda_i \boldsymbol{\xi}_i \quad (\text{and } \boldsymbol{\xi}_j^T A^T = \lambda_j \boldsymbol{\xi}_j^T)$$

Then

$$\begin{aligned} \boldsymbol{\mathcal{Z}}(\boldsymbol{\xi}_i \boldsymbol{\xi}_j^T) &= \gamma A \boldsymbol{\xi}_i \boldsymbol{\xi}_j^T A^T + A \boldsymbol{\xi}_i \boldsymbol{\xi}_j^T + \boldsymbol{\xi}_i \boldsymbol{\xi}_j^T A^T && \text{(E.0.4)} \\ &= \gamma \lambda_i \lambda_j \boldsymbol{\xi}_i \boldsymbol{\xi}_j^T + \lambda_i \boldsymbol{\xi}_i \boldsymbol{\xi}_j^T + \lambda_j \boldsymbol{\xi}_i \boldsymbol{\xi}_j^T && \text{(E.0.5)} \\ &= [\gamma \lambda_i \lambda_j + \lambda_i + \lambda_j] \boldsymbol{\xi}_i \boldsymbol{\xi}_j^T && \text{(E.0.6)} \end{aligned}$$

So $\boldsymbol{\xi}_i\boldsymbol{\xi}_j^T$ is an eigenvector of $\mathcal{Z}$ with eigenvalue $\gamma\lambda_i\lambda_j + \lambda_i + \lambda_j$. This motivates the following results.

Theorem E.1 *Let $\mathcal{Z}$ be the operator defined in (E.0.2). Let λ_i, λ_j be the distinct eigenvalues of A for $i, j = 1, 2, \ldots, m \leq n$. Then $\gamma\lambda_i\lambda_j + \lambda_i + \lambda_j$ is an eigenvalue of $\mathcal{Z}$. Conversely, let η_k, $k = 1, 2, \ldots, p \leq n^2$, be the distinct eigenvalues of $\mathcal{Z}$, then for each k,*

$$\eta_k = \gamma\lambda_i\lambda_j + \lambda_i + \lambda_j$$

for some i and some j.

Proof: Recognizing that $\mathcal{Z}(\mathbf{X})$ can be written as

$$\mathcal{Z}(\mathbf{X}) = \frac{1}{\gamma}[(\gamma\mathbf{A} + \mathbf{I}_n)\mathbf{X}(\gamma\mathbf{A} + \mathbf{I}_n)^T - \mathbf{X}] \tag{E.0.7}$$

which is just a transformation of the discrete-time Lyapunov equation operator, the result follows. See for example Theorem F2 in [16]. ▮

Corollary E.1 *Any matrix representation of the operator $\mathcal{Z}$ is nonsingular if and only if $\gamma\lambda_i\lambda_j + \lambda_i + \lambda_j \neq 0 \ \forall i, j$.*

Proof: Since the linear operator $\mathcal{Z}$ maps an n^2 dimensional linear space into itself, it has a matrix representation, i.e. one can write the n^2 equations corresponding to (E.0.1) as $\bar{\mathbf{A}}\bar{\mathbf{k}} = \bar{\mathbf{q}}$. The determinant of $\bar{\mathbf{A}}$ is the product of its eigenvalues, from which the result follows. ▮

Theorem E.2 *If $\gamma\lambda_i\lambda_j + \lambda_i + \lambda_j \neq 0 \ \forall i, j$, then. there exists a unique solution to equation (E.0.1). If, in addition, $|\gamma\lambda_i(\mathbf{A}) + 1| < 1$ then the solution can be written (i)*

$$\mathbf{X} = \gamma\kappa \sum_{k=0}^{\infty} (\gamma\mathbf{A} + \mathbf{I}_n)^k \mathbf{Q} (\gamma\mathbf{A}^T + \mathbf{I}_n)^k \tag{E.0.8}$$

or (ii)

$$\mathbf{X} = \mathbf{X}_0 + \gamma\mathbf{X}_1 + \gamma^2\mathbf{X}_2 + \ldots \tag{E.0.9}$$

where $\mathbf{X}_0$ *is the solution of the continuous-time Lyapunov equation*

$$\mathbf{A}\mathbf{X}_0 + \mathbf{X}_0\mathbf{A}^T + \gamma\kappa\mathbf{Q} = \mathbf{0} \tag{E.0.10}$$

and, for $n > 0$, $\mathbf{X}_n$ *is given by the solution of the continuous-time Lyapunov equation*

$$\mathbf{A}\mathbf{X}_n + \mathbf{X}_n\mathbf{A}^T + \mathbf{A}\mathbf{X}_{n-1}\mathbf{A}^T = \mathbf{0} \tag{E.0.11}$$

for $n \geq 2$.

Moreover, if $[\mathbf{A}, \mathbf{Q}^{1/2}]$ *is completely controllable then*

$$\begin{aligned}
\mathbf{X} = \mathbf{X}^T > \mathbf{0} \ &\textit{if} \ |\gamma\lambda_i(\mathbf{A}) + 1| < 1 \ \textit{and} \ \gamma\kappa > 0\\
\mathbf{X} = \mathbf{X}^T > \mathbf{0} \ &\textit{if} \ |\gamma\lambda_i(\mathbf{A}) + 1| > 1 \ \textit{and} \ \gamma\kappa < 0\\
\mathbf{X} = \mathbf{X}^T < \mathbf{0} \ &\textit{if} \ |\gamma\lambda_i(\mathbf{A}) + 1| < 1 \ \textit{and} \ \gamma\kappa < 0\\
\mathbf{X} = \mathbf{X}^T < \mathbf{0} \ &\textit{if} \ |\gamma\lambda_i(\mathbf{A}) + 1| > 1 \ \textit{and} \ \gamma\kappa > 0
\end{aligned}$$

Proof: Uniqueness follows from Corollary (E.1). Also using (E.0.7), equation (3.5.25) can be rewritten as

$$(\gamma\mathbf{A} + \mathbf{I}_n)\mathbf{X}(\gamma\mathbf{A} + \mathbf{I}_n)^T - \mathbf{X} + \gamma\kappa\mathbf{Q} = \mathbf{0} \tag{E.0.12}$$

By standard Lyapunov theory, the solution of (E.0.12) may be written in the form of (E.0.8) if $|\lambda_i(\gamma\mathbf{A} + \mathbf{I}_n)| < 1$, i.e. $|\gamma\lambda_i(\mathbf{A}) + 1| < 1$.

Part (ii) of the lemma follows via substitution.

It should be noted that complete controllability of $[\gamma\mathbf{A} + \mathbf{I}_n, \gamma\kappa\mathbf{Q}]$ is equivalent to complete controllability of $[\mathbf{A}, \mathbf{Q}^{1/2}]$. This can be shown using the PBH eigenvector test for controllability. $[\mathbf{A}, \mathbf{Q}^{1/2}]$ is controllable if and only if there is no left eigenvector of $\mathbf{A}$ orthogonal to $\mathbf{Q}^{1/2}$:

$$\mathbf{v}^T\mathbf{A} = \lambda\mathbf{v}^T \Leftrightarrow \mathbf{v}^T\mathbf{A} = \frac{\tilde{\lambda} - 1}{\gamma}\mathbf{w}^T \Leftrightarrow \gamma\mathbf{v}^T\mathbf{A} + \mathbf{v}^T = \tilde{\lambda}\mathbf{v}^T \Leftrightarrow \mathbf{v}^T(\gamma\mathbf{A} + \mathbf{I}_n) = \tilde{\lambda}\mathbf{v}^T$$

where $\tilde{\lambda} = \gamma\lambda + 1$. The positive/negative definiteness of the solution follows from Theorem 3.3 in [34]. If $\gamma\kappa > 0$ and $\pi(\gamma\mathbf{A} + \mathbf{I}_n) = 0$, then $\nu(\mathbf{X}) = 0 \ \Rightarrow \ \mathbf{X} > \mathbf{0}$. Also, if $\gamma\kappa > 0$ and $\nu(\gamma\mathbf{A} + \mathbf{I}_n) = 0$, then $\pi(\mathbf{X}) = 0 \ \Rightarrow \ \mathbf{X} < \mathbf{0}$. The cases $\gamma\kappa < 0$ follow naturally. ∎

The next result follows from equation (E.0.1) and continuity.

Lemma E.1 *From Theorem E.2, if* $|\gamma\lambda_i(\mathbf{A})+1|<1$ *with*

$$\mathbf{X}=\gamma\kappa\sum_{k=0}^{\infty}(\gamma\mathbf{A}+\mathbf{I}_n)^k\mathbf{Q}(\gamma\mathbf{A}^T+\mathbf{I}_n)^k$$

then

$$\lim_{\gamma\to 0}\mathbf{X}=\kappa\int_0^{\infty}e^{\mathbf{A}t}\mathbf{Q}e^{\mathbf{A}^Tt}\,dt \tag{E.0.13}$$

Lemma E.2 *Assume* κ *is independent of* γ. *If* $|\gamma\lambda_i(\mathbf{A})+1|-1$ *is of one sign for all* $i=1,\ldots,n$ *then the partial derivative of* $\mathbf{X}$ *with respect to* γ *satisfies*

$$\mathcal{D}_\gamma\mathbf{X}>\mathbf{0} \text{ if } \gamma\kappa>0 \tag{E.0.14}$$

$$\mathcal{D}_\gamma\mathbf{X}<\mathbf{0} \text{ if } \gamma\kappa<0 \tag{E.0.15}$$

Proof: Using equation (E.0.1) and taking partial derivatives with respect to γ gives

$$\gamma\mathbf{A}\mathcal{D}_\gamma\mathbf{X}\mathbf{A}^T+\mathbf{A}\mathcal{D}_\gamma\mathbf{X}+\mathcal{D}_\gamma\mathbf{X}\mathbf{A}^T+\mathbf{A}\mathbf{X}\mathbf{A}^T=\mathbf{0} \tag{E.0.16}$$

Using Theorem E.2, $\mathcal{D}_\gamma\mathbf{X}>\mathbf{0}$ if $|\gamma\lambda_i(\mathbf{A})+1|<1$ and $\mathbf{X}>\mathbf{0}$ (with $\mathbf{A}$ full rank). However if $\mathbf{X}>\mathbf{0}$, then this can only occur if $\gamma\kappa>0$. Alternatively $\mathcal{D}_\gamma\mathbf{X}>\mathbf{0}$ if $|\gamma\lambda_i(\mathbf{A})+1|>1$ and $\mathbf{X}<\mathbf{0}$. However if $\mathbf{X}<\mathbf{0}$, then this can only occur if $\gamma\kappa>0$. The first inequality is thus proved. The second inequality follows by similar arguments. ▮

Lemma E.3 *Assume* κ *is independent of* γ *and* $\gamma'\in[0,\gamma]$. *If* $|\gamma'\lambda_i(\mathbf{A})+1|-1$ *is of one sign for all* $i=1,\ldots,n$ *then the solution* $\mathbf{X}$ *to equation (E.0.1) satisfies*

$$\mathbf{X}>\mathbf{X}_0 \text{ if } \gamma\kappa>0 \tag{E.0.17}$$

$$\mathbf{X}<\mathbf{X}_0 \text{ if } \gamma\kappa<0 \tag{E.0.18}$$

where $\mathbf{X}_0$ *is the solution to equation (E.0.1) in the case* $\gamma=0$.

Proof: Using Taylor's theorem write

$$\mathbf{X} = \mathbf{X}_0 + \int_0^{\gamma} \mathcal{D}_\gamma \mathbf{X}(\gamma')\, d\gamma' \tag{E.0.19}$$

If $\mathcal{D}_\gamma \mathbf{X} > \mathbf{0}$ then $\mathbf{X} > \mathbf{X}_0$. From the previous lemma this occurs when $\gamma\kappa > 0$. Again, the second inequality is argued similarly. ▮

Appendix F

Algorithms Developed

This appendix lists the specific algorithms developed through the work of this thesis and outlines their function and use. The algorithms have been coded in in MATLAB,[1] and many require the Control System, Robust Control and Optimization Toolboxes for their use. A few of the algorithms require the SIMULINK package.

Figure F-1 shows all the programs and sub-programs developed along with the hierarchical structure of their operation. Note that the "starred" programs were not written by the author of this thesis.

F.1 Algorithms of Chapter 3

integrals.m This is the fundamental program used in Chapter 3. Its purpose is to implement the open-loop discretization algorithm. The program calls **systems.m** which contains all the essential parameters related to the examples of Section 3.9. The value of the parameter *EXAMPLE* determines which example is selected. The optimization is performed via the MATLAB optimization function **fminu**. The optimal discretized system is returned in the state space system $[F, G, H, J]$. A plot of the output of the signal invariant transformation and the optimal system is also produced (in response to the reference signal).

Generally the algorithm converges rapidly to the global minimum. However there are examples which are sensitive to the initial parameter selection. A few different

[1] MATLAB is a registered trademark of The MathWorks, Inc.

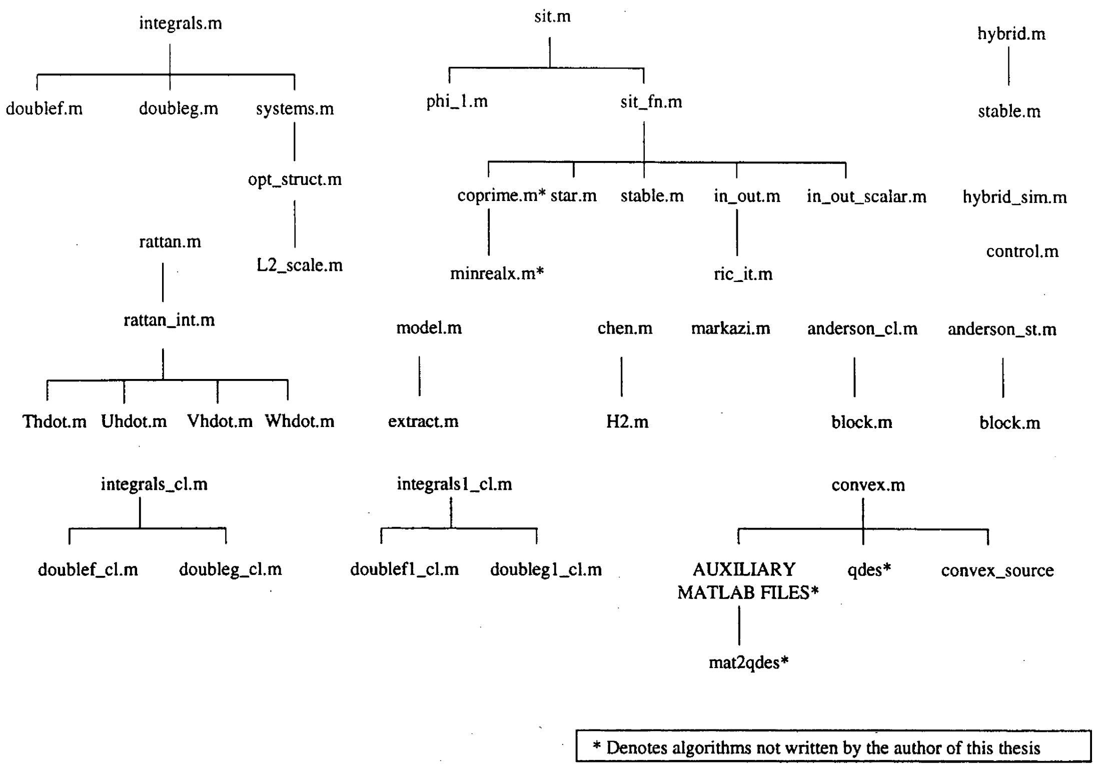

Figure F-1: Programs developed

starting points should be tried in these cases.

systems.m This file contains all the parameters related to the examples of Section 3.9. This includes a state space description of the system ($[ac, bc, cc, dc]$) and the signal model ($[ar, br, cr, dr]$). Other parameters are: the discretization period (T), the number of passes through the doubling algorithm (N), $\bar{\mathbf{A}}$ ($abar$), the integral approximation order (L), the options parameter required for the MATLAB optimization function **fminu** (options), the initial value of **p** (x), and finally the number of points to be plotted in the result (NN). The optimal discretized system is returned in the state space system $[F, G, H, J]$.

doublef.m This function is called by **integrals.m** through the MATLAB optimization function **fminu**. It evaluates the cost function (3.4.32) using the doubling algorithm for Lyapunov equations (c.f. Algorithm 3.1).

doubleg.m This function is also called by **integrals.m** through **fminu**. It evaluates the gradient of the cost function (3.4.32), again using the doubling algorithm.

opt_struct.m This function implements the theory of Section 3.9.3 to give a state space structure which minimizes the output noise gain, subject to a unity L_2 scaling constraint. The user must supply the integrator noise intensity $\mathbf{\Omega}_\eta$ (*omega_eta*), and the input noise intensity $\mathbf{\Omega}_u$ (*omega_u*).

L2_scale.m Given a diagonal matrix $\mathbf{D} \in \mathbb{R}^{n \times n}$, this function finds an orthogonal matrix $\mathbf{V}$ such that $(\mathbf{V}\mathbf{D}\mathbf{V}^T)_{jj} = \mathrm{tr}(\mathbf{D})/n \ \forall j$.

F.2 Algorithms of Chapter 4

sit.m This program implements Algorithm 4.1 for the H_2 optimal control method of controller discretization. The user must input the state space description of the plant ($[ap, bp, cp, dp]$), the controller ($[ac, bc, cc, dc]$) and the reference model ($[ar, br, cr, dr]$). Other parameters that must be selected are: the sampling period (T), the grid sizes ξ_1, ξ_2 corresponding to phase one and two of the algorithm respectively ($x1, x2$), and the time interval for the plots (t). The optimal discretized controller is returned in *S_C_star*.

phi_1.m This function is called from **sit.m** and evaluates the cost function Φ_1. The user must specify the grid size N.

sit_fn.m This function solves the primary optimization problem associated with the H_2 optimal control method of controller discretization. The user may have to adjust ($[ai, bi, ci, di]$) to ensure a continuous-time internal model (c.f. Section 4.2.3).

star.m This function finds the discrete-time adjoint of a system S.

stable.m This function decomposes a discrete-time system S into a sum of its stable and unstable parts. The user can specify a tolerance (*tol*) to distinguish stable poles from unstable poles.

in_out.m This function computes an inner-outer factorization of a system S in discrete-time. It is based on the theory of [36]. The user can include an optional second argument to specify that the Riccati equations are to be solved by iteration using **ric_it.m**.

in_out_scalar.m This function computes an inner-outer factorization for a SISO system S. The algorithm simply reflects poles and zeros across the unit circle.

ric_it.m A function for solving the discrete-time Riccati equation by iteration. The user specifies the number of iterations in the function call.

convex.m This program implements the convex optimization algorithm for controller discretization. The user must specify the state space descriptions of the plant ($[apc, bpc, cpc, dpc]$), the controller ($[ac, bc, cc, dc]$), and that of the reference system ($[ar, br, cr, dr]$). The sampling period (T), the number of intersample divisions (N), and *n_sample* are also needed.

qdes This is an executable C program called from **convex.m** for solving the convex optimization problem. This version is compatible with a sun4 computer.

convex_source This is the source file called from **convex.m** and used by **qdes**. The user must specify the number of exogenous inputs (*n_exog*), the number of regulated outputs (*n_reg*), the number of sensed outputs (*n_sens*), the number of actuator inputs (*n_act*), and the variables *n_sample*, and *n_tap* that appear in Section 4.3.3. The norms used in the optimization and the constraints must also be input.

AUXILIARY MATLAB FILES This range of functions not written by the author of this thesis includes **fir2ss.m, loadq.m, q2comp.m, gethzw.m, kl2ttt.m, kl2ttt_fn.m**, and **midimpulse.m**. These functions are called from **convex.m** and are used to process the numerical data to and from **qdes**. These functions are included in the qdes distribution file. A full description of these function is available from the distribution site.

mat2qdes This is an auxiliary C program used by **qdes** to write impulse response data to disk files.

minrealx.m This is a modified version of the MATLAB function **minreal.** It has a high tolerance value for determining unobservable and uncontrollable states of a system.

coprime.m This function returns state space realizations of elements of a doubly coprime factorization of a system $[A, B, C, D]$ based on a state feedback gain $Ksfb$ and an observer gain $Lobs$ ($A - B \times Ksfb$ and $A - Lobs \times C$ must be stable).

integrals_cl.m This program implements the integral approximation method of controller discretization. The user must give the state space description of the plant ($[ap, bp, cp, dp]$), the controller ($[ac, bc, cc, dc]$), and the reference model ($[ar, br, cr, dr]$). A number of parameters need to be specified: the sampling period (T), the number of passes through the doubling algorithm (N), $\bar{\mathbf{A}}$ (*abar*), the order of the integral approximations (L), the number of intersample divisions (*division*), the initial parameter **p** (x), the upper and lower bounds on the parameter region (vlb, vub), the number of intersample divisions (*division*), and the options for the MATLAB function **minimax** (*options*). The state space description of the discretized controller is returned in $[F, G, H, J]$.

Generally the algorithm should be run two or three times, first with a small value of N (typically 2 or 3) and then with a larger value. The optimal **p** found at each step should be used as the value of x in the next iteration.

doublef_cl.m This function evaluates the cost function (4.4.1). This is called by **integrals_cl.m** through **minimax.**

doubleg_cl.m This function computes the gradient of (4.4.1). This is again called by **integrals_cl.m** through **minimax.**

integrals1_cl.m, doublef1_cl.m, doubleg1_cl.m These are equivalent functions to **integrals_cl.m, doublef_cl.m,** and **doubleg_cl.m** except that the integral approximations are restricted to first order, which alleviates some of the numerical problems of the general algorithms. Obviously the parameter L is not required.

hybrid.m This function performs a sampled-data controller reduction using the algorithm of [56]. The function has input parameters comprised of: the continuous-time plant (S), the digital controller (S_C), the fast sampling factor (N), and the sampling period (T). An optional parameter (*cr_order*) which specifies the order of the reduced controller can be included. If this is not present, the order is asked for at a later point. The outputs are the reduced order controller (S_CR), the discrete-time closed-loop system with reduced controller (S_CLR), and the weighted Hankel singular values (HSV).

F.3 Algorithms of Chapter 5

model.m From experimental data (contained in file **identification.dat**), this program finds the parameters ($\alpha_1, \alpha_2, \alpha_3, \beta_1, \gamma_1, \gamma_2$) corresponding to the non-linear model of the two-tank apparatus (equations (5.2.4) and (5.2.5)). It also generates a linearized model of the system in state space form ($[A, B, C, D]$).

extract.m This function is called by **model.m** and is used to extract data from the file **identification.dat**.

control.m This program is used for the design of the analog LQG controller for use on the two-tank apparatus. In order to reject constant disturbances at the output, the design includes an integrator.

rattan.m This program implements Rattan's method of controller discretization. The designer must specify the state space descriptions of the plant ($[A, B, C, D]$) and the controller ($[Ac, Bc, Cc, Dc]$) must be supplied by the user. Furthermore the sampling period (T) and the range of integration ($gamma\_1, gamma\_2$) must be given. Options related to the step sizes used in the Adams integration method can also be included. The digital controller is returned in $[Acd, Bcd, Ccd, Dcd]$.

Thdot.m, Uhdot.m, Vhdot.m, Whdot.m These are functions called by **rattan.m** to calculate the integrands of the four integrals associated with Rattan's method.

rattan_int.m This is a block generated by SIMULINK which is called through **rattan.m**. It sets up the call to the four previous functions.

anderson_cl.m,anderson_st.m These programs implement Anderson and Keller's closed-loop method and stability method of controller discretization. The user must specify state space descriptions of the plant ($[ap, bp, cp, dp]$), and the analog controller ($[ac, bc, cc, dc]$). A weighting function ($[aw, bw, cw, dw]$) must also be specified in the closed-loop method. The sampling period (T) and the fast sampling factor (N) must be specified. The resulting sampled-data controller is returned in $[acp, bcp, ccp, dcp]$.

block.m This is a function called from **anderson_cl.m** and **anderson_st.m** to perform the "blocking".

markazi.m This program implements Markazi and Hori's method of controller discretization in the case when the plant is stable. The user must specify transfer function descriptions of the plant ($nump, denp$) and the analog controller ($numc, denc$). The digital controller is returned in $[numcd, dencd]$.

chen.m This program implements Chen's direct method of H_2 sampled-data controller design for the problem of controller discretization. The user must specify state space descriptions of the plant ($[ap, bp, cp, dp]$), the analog controller ($[ac, bc, cc, dc)$, and a weighting function ($[aw, bw, cw, dw]$). The sampling period (T) and the fictitious noise source intensity (rho) must also be specified. The optimal digital controller is returned in $[acp, bcp, ccp, dcp]$).

H2.m This function is called from **chen.m** which solves the discrete-time H_2 problem. The function utilizes the discrete-time Kalman filter (rather than predictor).

hybrid_sim.m This is a SIMULINK block which enables the comparison of an analog closed-loop system and a hybrid closed-loop system. The following must be entered in the MATLAB workspace: the continuous-time plant ($[A, B, C, D]$), the continuous-time controller ($[Acd, Bcd, Ccd, Dcd]$), the sampled-data controller ($[Ac, Bc, Cc, Dc]$), and the sampling period (T).

Bibliography

[1] B.D.O. Anderson. Optimizing the discretization of continuous-time controllers. In H. Kimura and S. Kodama, editors, *Proceedings of the international symposium on mathematical theory of networks and systems MTNS-91*, pages 475–480, Japan, 1992. Mita press.

[2] B.D.O Anderson and Moore. *Optimal Filtering.* Prentice-Hall, Englewood Cliffs, New Jersey, 1979.

[3] B.D.O Anderson and J.B. Moore. *Optimal Control - Linear Quadratic Methods.* Prentice Hall, Englewood Cliffs, NJ, 1990.

[4] A. Antoniou. *Digital filters: Analysis and design.* McGraw Hill, New York, 1979.

[5] K.J. Aström, P. Hagander, and J. Sternby. Zeros of sampled systems. *Automatica*, 20(1):31–38, 1984.

[6] K.J. Aström and B. Wittenmark. *Computer Controlled Systems: Theory and Design.* Prentice Hall, Englewood Cliffs, NJ, 1984.

[7] J.A. Ball, I. Gohberg, and L. Rodman. *Interpolation of Rational Matrix Functions*, volume OT 45 of *Operator Theory: Advances and Applications.* Birkhäuser Verlag, Basel Boston Berlin, 1990.

[8] J.A. Ball, I. Gohberg, and L. Rodman. Interpolation problems for rational matrix function and system theory. In H. Kimura and S. Kodama, editors, *Proceedings of the international symposium on mathematical theory of networks and systems MTNS-91*, pages 3–12, Japan, 1992. Mita press.

[9] B. Bamieh and J.B. Pearson. The H_2 problem for sampled-data systems. Technical Report 9104, Department of Electrical and Computer Engineering, Rice University, 1991.

[10] B. Bamieh and J.B. Pearson. A general framework for linear periodic systems with application to H_∞ sampled-data systems. *IEEE Transactions on Automatic Control*, AC-37:418–435, 1992.

[11] K.G. Beauchamp. *Signal Processing using analog and digital techniques.* Allen and Unwin, London, 1973.

[12] A. Ben-Zwi and M. Preiszler. Comparison of discretization methods. Technical report, Rafael, Israel MOD, 1979. in Hebrew.

[13] S.P. Boyd, V. Balakrishnan, C.H. Barratt, N.M. Khraishi, X. Li, D.G. Meyer, and S. Norman. A new CAD method and associated architectures for linear controllers. *IEEE Transactions on Automatic Control*, AC-33:268–282, 1988.

[14] S.P. Boyd and C.H. Barratt. *Linear Controller Design - Limits of Performance.* Prentice-Hall, Englewood Cliffs, NJ, 1991.

[15] L.T. Bruton and D.A. Vaughan-Pope. Synthesis of digital ladder filters from LC filters. *IEEE Transactions on Circuits and Systems*, CAS 23(6):395–402, 1976.

[16] C.T. Chen. *Analysis and Synthesis of Linear Control Systems.* Holt, Rinehart and Winston, New York, 1975.

[17] T. Chen. A simple derivation of the H_2-optimal sampled-data controllers. *Systems and Control Letters*, 20(1):49–56, 1993.

[18] T. Chen and B.A. Francis. On the L_2-induced norm of a sampled-data system. *Systems and Control Letters*, 15:211–219, 1990.

[19] T. Chen and B.A. Francis. Input-output stability of sampled-data systems. *IEEE Transactions on Automatic Control*, AC-36(1):50–58, 1991.

[20] T. Chen and B.A. Francis. H_2-optimal sampled-data control. *IEEE Transactions on Automatic Control*, AC-36(1):387–397, 1991.

[21] T. Chen and B.A. Francis. Stability of sampled-data feedback systems. *IEEE Transactions on Automatic Control*, AC-36:50–58, 1991.

[22] A.J.O. Cruickshank. Time series and z-transform methods of analysis of linear and nonlinear control systems. In *Proceedings of 1st International IFAC Congress*, pages 277–285, Moscow, 1960.

[23] J. Daniel. *Approximate Minimization of Functionals.* Prentice-Hall, 1971.

[24] A.G. Deczky. Synthesis of recursive digital filters using the minimum l_p-error criterion. *IEEE Transactions on Audio Electroacoustics*, AU-20:257–263, 1972.

[25] C.A. Desoer, R.W. Liu, J. Murray, and R. Saeks. Feedback system design: The fractional representation approach to analysis and synthesis. *IEEE Transactions on Automatic Control*, AC-25:399–412, 1980.

[26] G.E. Dullerud and B.A Francis. L_1 analysis and design of sampled-data systems. *IEEE Transactions on Automatic Control*, AC-37(4):436–446, 1992.

[27] D. Enns. *Model Reduction for Control Design.* PhD thesis, Stanford University, Stanford, California, 1984.

[28] W. Findeison. Automatic control techniques. Technical report, Warsaw: PWN, 1978. in Polish.

[29] B.A. Francis and T.T. Georgiou. Stability theory for linear time-invariant plants with periodic digital controllers. *IEEE Transactions on Automatic Control*, AC-33(9):820–832, 1988.

[30] B.A. Francis and W.M. Wonham. The internal model principle of control theory. *Automatica*, 12:457–465, 1976.

[31] G.F. Franklin and A. Emami-Naeini. Design of ripple-free multivariable robust servomechanisms. *IEEE Transactions on Automatic Control*, AC-31:661–664, 1986.

[32] G.F. Franklin and J.D. Powell. *Digital Control of Dynamic System.* Addison-Wesley Publishing Company, 1980.

[33] R. Gessing. Comments about frequency response plots of Keller and Anderson and Rattan. *IEEE Transactions on Automatic Control*, AC-39:1770–1771, 1994.

[34] K. Glover. All optimal hankel-norm approximations of linear multivariable systems and their L_∞-error bounds. *International Journal of Control*, 39:1115–1193, 1984.

[35] B. Gold and C.M Rader. *Digital processing of signals.* McGraw-Hill, New York, 1969.

[36] D.W. Gu, M.C. Tsai, S.D. O'Young, and I. Postlethwaite. State-space formulae for discrete-time H^∞ optimization. *International Journal of Control*, 49:1683–1723, 1989.

[37] S.R. Hall. Comments on two methods for designing a digital equivalent to a continuous control system. *IEEE Transactions on Automatic Control*, AC-39(2):420–421, 1994.

[38] S. Hara and P.T. Kabamba. On optimizing the induced norm of a sampled-data system. In *Proceedings of the Conference on Decision and Control*, 1990.

[39] E.L. Jury. *Theory and application of the z-transform method.* Wiley, New York, 1964.

[40] P.T. Kabamba and S. Hara. On computing the induced norm of a sampled-data system. In *Proceedings of the American Control Conference*, 1990.

[41] P.T. Kabamba and S. Hara. Worst case analysis and design of sampled data control systems. *IEEE Transactions on Automatic Control*, AC-38:214–223, 1993.

[42] P. Katz. *Digital Control Using Microprocessors.* Prentice hall, Englewood Cliffs, NJ, 1981.

[43] J.P. Keller and B.D.O. Anderson. A new approach to the discretization of continuous-time controllers. In *Proceedings of the American Control Conference*, pages 1127–1132, San Diego, 1990.

[44] S.G. Kellison. *Fundamentals of Numerical Analysis.* Richard D. Irwin, Inc., 1975.

[45] R.A. Kennedy and R.J. Evans. Digital resign on a continuous controller based on closed loop performance. In *Proceedings of the 29th IEEE Conference on Decision and Control*, Hawaii, 1990.

[46] P.P. Khargonekar and N. Sivashankar. H_2 optimal control for sampled-data systems. *Systems and Control Letters*, 17(6):425–436, 1991.

[47] Z. Kowalczuk. Digital modelling of continuous systems. PID control. In *Proceedings of 1st IASTED Conference on Applied Modelling and Simulation*, pages 21–28, Lyon, France, 1981.

[48] Z. Kowalczuk. Certain relations of normal methods of transformations. *Systems Science*, 2:5–29, 1983.

[49] Z. Kowalczuk. A normal triangle and normal methods of transformations in the discrete approximation of continuous system. *Arc. Autom. Telemech*, 28(1-2):15–32, 1983.

[50] Z. Kowalczuk. Discrete approximation of continuous-time systems: a survey. *IEE Proceedings-G*, 140(4), 1993.

[51] E. Kreyszig. *Advanced Engineering Mathematic.* John Wiley and Sons, 1983.

[52] B.C. Kuo. *Digital Control Systems.* Saunders College Publishing, 1992.

[53] B.C. Kuo and D.W. Peterson. Optimal discretization of continuous-data control systems. *Automatica*, 9:125–129, 1973.

[54] B.C. Kuo, G. Singh, and R. Yackel. Digital approximation of continuous-data control systems by point-by-point state comparison. *Computers and Electrical Engineering*, 1:155–170, 1973.

[55] E. Levitin and B. Polyak. Constrained minimization methods. *USSR Computational Math. and Math. Physics*, 6(5):1–50, 1966.

[56] A.G. Madievski and B.D.O. Anderson. Sampled-data controller reduction procedure. submitted to *IEEE Transactions on Automatic Control.*

[57] M. Marden. The geometry of the zeros of a polynomial in a complex variable. In *Mathematical surveys: Functions of complex variables.* American Mathematical Society, 1949.

[58] A.H.D. Markazi and N. Hori. A new method with guaranteed stability for discretization of continuous-time control systems. In *American Control Conference*, pages 1397–1401, 1992.

[59] D.F. Miller. Multivariable linear digital control via state-space output matching. *Opt. Control Applic. Meth*, 6:13, 1985.

[60] H.H. Niemann and O. Jannerup. A frequency domain design method for sampled data compensators. In *Proceedings American Control Conference*, pages 1156–1158, Arizona, 1990.

[61] K. Nordström. Optimal input matching in sampled data control systems. In *Proceedings of the 31st Conference on Decision and Control*, pages 1941–1943, Arizona, 1992.

[62] A.V. Oppenheim and R.W. Schafer. *Digital signal processing.* Prentice Hall, Englewood Cliffs, 1975.

[63] L.R. Rabiner and B. Gold. *Theory and application of digital signal processing.* Prentice Hall, Englewood Cliffs, 1975.

[64] C.M. Radar and B. Gold. Digital filter design techniques in the frequency domain. In *Proceedings IEEE*, pages 149–171, 55 1967.

[65] K.S. Rattan. Digitizing of existing continuous control systems. *IEEE Transactions on Automatic Control*, AC-29:282–285, 1984.

[66] K.S. Rattan. Compensating for computational delay in digital equivalent of continuous control systems. *IEEE Transactions on Automatic Control*, AC-34(8):895–899, 1989.

[67] L.S. Shieh, Y.F. Chang, and R.E. Yates. Model simplification and digital design of multivariable sampled-data control systems via a dominant-data matching method. In *Proceedings IFAC Symp. Theory Applic. Digital Control*, New Delhi, 1982.

[68] G. Singh, B.C. Kuo, and R.A. Yackel. Digital approximation by point-by-point state matching with higher-order holds. *International Journal of Control*, 20:81–90, 1974.

[69] J.M. Smith. *Mathematical modelling and digital simulation for scientists and engineers.* Wiley, New York, 1977.

[70] S.D. Stearns and D.R. Hush. *Digital signal analysis.* Prentice Hall, Englewood Cliffs, 1990.

[71] D. Tabak. Digitalisation of control systems. *Computer Aided Design*, 3:13–18, 1971.

[72] G.B. Thomas and R.L. Finney. *Calculus and Analytic Geometry.* Addison-Wesley Publishing Company, 1979.

[73] P.M. Thompson, R.L. Dailey, and J.C. Doyle. New conic sectors for sampled-data feedback systems. *Systems and control letters*, 7:395–404, 1986.

[74] P.M. Thompson, G. Stein, and M. Athans. Conic sectors for sampled-data feedback systems. *Systems and control letters*, 3:77–82, 1983.

[75] H.T. Toivonen. Sampled-data control of continuous-time systems with an H_∞ optimality criterion. *Automatica*, 28:45–54, 1992.

[76] A. Tustin. A method of analysing the behaviour of linear systems in terms of time series. In *Journal IEE*, pages 130–142, 94 1947.

[77] S. Urikura and A. Nagata. Ripple-free deadbeat control for sampled-data systems. *IEEE Transactions on Automatic Control*, AC-32(6):474–482, 1987.

[78] R. Vich. Selective properties of digital filters obtained by convolution approximation. *Electronic Letters*, 4(1):1–2, 1968.

[79] R. Vich. Approximation in digital filter synthesis based on time response invariance. *Electronic Letters*, 6(14):442–445, 1970.

[80] R. Vich. Properties of digital bandpasses obtained by convolution approximation for zero frequency. *Electronic Letters*, 6(14):440–442, 1970.

[81] R. Vich. Two methods for the construction of transfer functions of digital integrators. *Electronic Letters*, 7(15):442–445, 1971.

[82] R. Vich. *$\mathcal{Z}$ transform theory and applications.* D. Reidel, Dordrecht, 1987.

[83] M. Vidyasagar. *Control System Synthesis: A Factorization Approach.* M.I.T. Press, Cambridge, MA, 1985.

[84] D. Williamson. *Digital Control and Implementation: Finite Wordlength Considerations.* Prentice Hall International, 1991.

[85] R.A. Yackel, B.C. Kuo, and G. Singh. Digital redesign of continuous systems by matching of states at multiple sampling periods. *Automatica*, 10:105–111, 1974.

[86] Y. Yamamoto. A new approach to sampled-data control systems-a function space approach method. In *Proceedings of the Conference on Decision and Control*, 1990.

[87] D.C. Youla, J.J. Bongiorno Jr., and H.A. Jabr. Modern Wiener-Hopf design of optimal controllers part I. *IEEE Transactions on Automatic Control*, AC-21(1):3–14, 1976.

[88] D.C. Youla, J.J. Bongiorno Jr., and H.A. Jabr. Modern Wiener-Hopf design of optimal controllers part II. *IEEE Transactions on Automatic Control*, AC-21(6):319–330, 1976.